WEBIST 2014

10th International Conference on Web Information Systems and Technologies

PROCEEDINGS

Volume 1

Barcelona, Spain

3 - 5 April, 2014

Sponsored by:

INSTICC

Technically sponsored by:

ERCIS

WEBIST 2014

Proceedings of the
10th International Conference on
Web Information Systems and Technologies

Volume 1

Barcelona, Spain

3 - 5 April, 2014

Sponsored by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

Technically Sponsored by
ERCIS – European Research Center for Information Systems

Copyright © 2014 SCITEPRESS – Science and Technology Publications
All rights reserved

Edited by Valérie Monfort and Karl-Heinz Krempels

Printed in Portugal
ISBN: 978-989-758-023-9
Depósito Legal: 372333/14

http://www.webist.org
webist.secretariat@insticc.org

BRIEF CONTENTS

INVITED SPEAKERS IV

ORGANIZING AND STEERING COMMITTEES V

PROGRAM COMMITTEE VI

AUXILIARY REVIEWERS X

SELECTED PAPERS BOOK X

FOREWORD XI

CONTENTS XIII

INVITED SPEAKERS

Steven Willmott

3scale

Spain

Fabien Gandon

INRIA

France

Andreas Pfeiffer

Hubject GmbH - joint venture of BMW Group, Bosch, Daimler, EnBW, RWE & Siemens

Germany

Zakaria Maamar

Zayed University

U.A.E.

ORGANIZING AND STEERING COMMITTEES

CONFERENCE CHAIR

Valérie Monfort, Université de Paris1 Panthéon Sorbonne, France

PROGRAM CHAIR

Karl-Heinz Krempels, RWTH Aachen University, Germany

PROCEEDINGS PRODUCTION

Marina Carvalho, INSTICC, Portugal

Helder Coelhas, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Ana Guerreiro, INSTICC, Portugal

Andreia Moita, INSTICC, Portugal

Raquel Pedrosa, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

Cláudia Pinto, INSTICC, Portugal

Sara Santiago, INSTICC, Portugal

Fábio Santos, INSTICC, Portugal

José Varela, INSTICC, Portugal

CD-ROM PRODUCTION

Pedro Varela, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

André Lista, INSTICC, Portugal

Mara Silva, INSTICC, Portugal

SECRETARIAT

Carla Mota, INSTICC, Portugal

WEBMASTER

Susana Ribeiro, INSTICC, Portugal

PROGRAM COMMITTEE

Jose Luis Herrero Agustin, University of Extremadura, Spain

Mugurel Ionut Andreica, Polytechnic University of Bucharest, Romania

Guglielmo de Angelis, CNR - IASI, Italy

Margherita Antona, Foundation for Research and Technology - Hellas (FORTH), Greece

Valeria De Antonellis, University of Brescia, Italy

Liliana Ardissono, Universita' Di Torino, Italy

Giuliano Armano, University of Cagliari, Italy

Ismailcem Budak Arpinar, University of Georgia, U.S.A.

Elarbi Badidi, United Arab Emirates University, U.A.E.

Andrea Ballatore, University College Dublin, Ireland

David Bell, Brunel University, U.K.

Orlando Belo, University of Minho, Portugal

Werner Beuschel, CRADL Lab, Germany

Christoph Bussler, VoxeoLabs, Inc., U.S.A.

Maria Claudia Buzzi, CNR, Italy

Elena Calude, Massey University, Institute of Natural and Mathematical Sciences, New Zealand

Pasquina Campanella, University of Bari "Aldo Moro", Italy

Cinzia Cappiello, Politecnico di Milano, Italy

Sven Casteleyn, Universitat Jaume I, Spain

Federica Cena, University of Torino, Italy

Weiqin Chen, University of Bergen, Norway

Dickson Chiu, Dickson Computer Systems, Hong Kong

Chin-Wan Chung, Korea Advanced Institute of Science and Technology (KAIST), Korea, Republic of

Christophe Claramunt, Naval Academy Research Institute, France

Mihaela Cocea, University of Portsmouth, U.K.

Martine De Cock, Ghent University, Belgium

Christine Collet, Grenoble Institute of Technology, France

Marco Comuzzi, City University London, U.K.

Isabelle Comyn-Wattiau, Cnam & Essec, France

Anna Corazza, University of Naples "Federico II", Italy

Daniel Cunliffe, University of South Wales, U.K.

Florian Daniel, University of Trento, Italy

Mats Daniels, Uppsala University, Sweden

Mark M. Davydov, Independent Consultant, U.S.A.

Steven Demurjian, University of Connecticut, U.S.A.

Enrico Denti, Alma Mater Studiorum - Università di Bologna, Italy

Stefan Dessloch, Kaiserslautern University of Technology, Germany

Oscar Díaz, University of Basque Country, Spain

Josep Domingo-Ferrer, Universitat Rovira i Virgili, Spain

Atilla Elci, Aksaray University, Turkey

Vadim Ermolayev, Zaporozhye National University, Ukraine

Larbi Esmahi, Athabasca University, Canada

Davide Eynard, University of Lugano, Switzerland

Alexander Felfernig, Technische Universität Graz, Austria

Anna Fensel, STI Innsbruck, University of Innsbruck, Austria

Miriam Fernandez, The Open University, U.K.

Joao Carlos Amaro Ferreira, ISEL, Portugal

Josep-Lluis Ferrer-Gomila, Balearic Islands University, Spain

Karla Donato Fook, IFMA - Instituto Federal de Educação Ciência e Tecnologia do Maranhão, Brazil

Geoffrey Charles Fox, Indiana University, U.S.A.

PROGRAM COMMITTEE (CONT.)

Pasi Fränti, Speech and Image Processing Unit, University of Eastern Finland, Finland

Britta Fuchs, FH Aachen, Germany

Giovanni Fulantelli, Italian National Research Council, Italy

Ombretta Gaggi, Università di Padova, Italy

John Garofalakis, University of Patras, Greece

Panagiotis Germanakos, University of Cyprus, Cyprus

Massimiliano Giacomin, Università degli Studi di Brescia, Italy

José Antonio Gil, Universitat Politècnica de València, Spain

Anna Goy, University of Torino, Italy

Thomas Greene, M.I.T., U.S.A.

Ratvinder Grewal, Laurentian University, Canada

Begoña Gros, University of Barcelona, Spain

William Grosky, University of Michigan - Dearborn, U.S.A.

Angela Guercio, Kent State University, U.S.A.

Francesco Guerra, University of Modena and Reggio Emilia, Italy

Miguel Guinalíu, Universidad de Zaragoza, Spain

Shanmugasundaram Hariharan, TRP Engineering College, India

Ioannis Hatzilygeroudis, University of Patras, Greece

Stylianos Hatzipanagos, King's College London, U.K.

A. Henten, Aalborg University, Denmark

Emilio Insfran, Universitat Politècnica de València, Spain

Ivan Ivanov, SUNY Empire State College, U.S.A.

Kai Jakobs, RWTH Aachen University, Germany

Dietmar Jannach, Technical University of Dortmund, Germany

Monique Janneck, Luebeck University of Applied Sciences, Germany

Ivan Jelinek, Czech Technical University in Prague, Czech Republic

Yuh-Jzer Joung, National Taiwan University, Taiwan

Carlos Juiz, Universitat de les Illes Balears, Spain

Katerina Kabassi, Tei of the Ionian Islands, Greece

Georgia Kapitsaki, University of Cyprus, Cyprus

George Karabatis, Umbc, U.S.A.

Sokratis Katsikas, University of Piraeus, Greece

Natalya Keberle, Zaporozhye National University, Ukraine

Matthias Klusch, Deutsches Forschungszentrum für Künstliche Intelligenz, Germany

In-Young Ko, Korea Advanced Institute of Science and Technology, Korea, Republic of

Waldemar W. Koczkodaj, Laurentian University, Canada

Hiroshi Koide, Kyushu Institute of Technology, Japan

Fotis Kokkoras, TEI of Thessaly, Greece

Agnes Koschmider, KIT - U, Germany

Karl-Heinz Krempels, RWTH Aachen University, Germany

Tsvi Kuflik, The University of Haifa, Israel

Peep Kungas, University of Tartu, Estonia

Daniel Lemire, TELUQ, Canada

Stefania Leone, ETH Zürich, U.S.A.

Kin Fun Li, University of Victoria, Canada

Weigang Li, University of Brasilia, Brazil

Xitong Li, MIT Sloan School of Management, U.S.A.

Dongxi Liu, CSIRO, Australia

Xiaozhong Liu, Indiana University, U.S.A.

Xumin Liu, Rochester Institute of Technology, U.S.A.

Ying Liu, Cardiff University, U.K.

PROGRAM COMMITTEE (CONT.)

Leszek Maciaszek, Wroclaw University of Economics, Poland and Macquarie University, Sydney, Australia

Cristiano Maciel, Universidade Federal de Mato Grosso, Brazil

Michael Mackay, Liverpool John Moores University, U.K.

Tim A. Majchrzak, University of Münster, Germany

Dwight Makaroff, University of Saskatchewan, Canada

Massimo Marchiori, University of Padua, Italy

Kazutaka Maruyama, Meisei University, Japan

Michael Melliar-Smith, University of California, U.S.A.

Tarek Melliti, University of Evry, France

Ingo Melzer, Daimler AG, Germany

Emilia Mendes, Blekinge Institute of Technology, Sweden

Panagiotis Metaxas, Wellesley College, U.S.A.

Abdelkrim Meziane, CERIST Alger, Algeria

Tommi Mikkonen, Institute of Software Systems, Tampere University of Technology, Finland

Valérie Monfort, Université de Paris1 Panthéon Sorbonne, France

Louise Moser, University of California, Santa Barbara, U.S.A.

Tomasz Muldner, Acadia University, Canada

Stavros Nikolopoulos, University of Ioannina, Greece

Vit Novacek, Digital Enterprise Research Institute, Nuig, Ireland

Jeff Z. Pan, University of Aberdeen, U.K.

Laura Papaleo, Province of Genova, Italy

Kyparisia Papanikolaou, ASPETE, Greece

Marcin Paprzycki, Polish Academy of Sciences, Poland

Eric Pardede, La Trobe University, Australia

Kalpdrum Passi, Laurentian University, Canada

Viviana Patti, University of Torino, Italy

David Paul, The University of Newcastle, Australia

Toon De Pessemier, Ghent University - iMinds, Belgium

Alfonso Pierantonio, University of L'Aquila, Italy

Luis Ferreira Pires, University of Twente, The Netherlands

Pierluigi Plebani, Politecnico Di Milano, Italy

Jim Prentzas, Democritus University of Thrace, Greece

Birgit Pröll, Johannes Kepler University Linz, Austria

Dana Al Qudah, University of Warwick, U.K.

Werner Retschitzegger, Johannes Kepler University, Austria

Thomas Risse, L3S Research Center, Germany

Thomas Ritz, FH Aachen, Germany

Davide Rossi, University of Bologna, Italy

Gustavo Rossi, Lifia, Argentina

Davide Di Ruscio, University of L'Aquila, Italy

Maytham Safar, Kuwait University, Kuwait

Aziz Salah, Université du Québec à Montréal, Canada

Yacine Sam, University Tours, France

Comai Sara, Politecnico di Milano, Italy

Anthony Savidis, Institute of Computer Science, FORTH, Greece

Bernhard Schandl, Gnowsis.com, Austria

Claudio Schifanella, Università degli Studi di Torino, Italy

Harald Schöning, Software AG, Germany

Hamida Seba, University Lyon 1, France

Jochen Seitz, Technische Universität Ilmenau, Germany

Marianna Sigala, University of the Aegean, Greece

PROGRAM COMMITTEE (CONT.)

Marten van Sinderen, University of Twente, The Netherlands

Richard Soley, Object Management Group, Inc., U.S.A.

Anna Stavrianou, Xerox Research Centre Europe, France

Nenad Stojanovic, FZI at the University of Karlsruhe, Germany

Hussein Suleman, University of Cape Town, South Africa

Christoph Terwelp, RWTH Aachen University, Germany

Dirk Thissen, RWTH Aachen University, Germany

Thanassis Tiropanis, University of Southampton, U.K.

Giovanni Toffetti, IBM Research Lab Haifa, Israel

Riccardo Torlone, Università Roma Tre, Italy

Guy Tremblay, Universite du Quebec a Montreal, Canada

Th. Tsiatsos, Department of Informatics, Aristotle University of Thessaloniki, Greece

George Tsihrintzis, University of Piraeus, Greece

Athina Vakali, Aristotle University, Greece

Geert Vanderhulst, Alcatel-Lucent Bell Labs, Belgium

Jari Veijalainen, University of Jyvaskyla, Finland

Maria Esther Vidal, Universidad Simon Bolivar, Venezuela

Maria Virvou, University of Piraeus, Greece

Petri Vuorimaa, Aalto University, Finland

Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn Malaysia, Malaysia

Fan Wang, Microsoft, U.S.A.

Jason Whalley, Strathclyde University, U.K.

Bebo White, Stanford University, U.S.A.

Maarten Wijnants, Hasselt University, Belgium

Manuel Wimmer, Technische Universität Wien, Austria

Marco Winckler, University Paul Sabatier (Toulouse 3), France

Viacheslav Wolfengagen, Institute JurInfoR, Russian Federation

Bin Xu, Tsinghua University, China

Guandong Xu, University of Technology Sydney, Australia

Amal Zouaq, Royal Military College of Canada, Canada

AUXILIARY REVIEWERS

Alberto De La Rosa Algarin, University of Connecticut, U.S.A.

Markus C. Beutel, RWTH Aachen University, Germany

Matteo Ciman, University of Padua, Italy

José Cordeiro, Polytechnic Institute of Setúbal / INSTICC, Portugal

Javier Espinosa, LAFMIA lab, France

Golnoosh Farnadi, Ghent University, The Netherlands

Sevket Gökay, RWTH Aachen University, Germany

Beatriz Gomez, University of the Balearic Islands, Spain

Nuno Pina Gonçalves, EST-Setúbal / IPS, Portugal

Wolfgang Kluth, RWTH Aachen University, Germany

Fangfang Li, AAi, Australia

José António Sena Pereira, IPS - ESTSetúbal, Portugal

Laura Po, University of Modena and Reggio Emilia, Italy

Michael Rogger, STI, Austria

Christian Samsel, RWTH Aachen University, Germany

Alexander Semenov, University of Jyvaskyla, NRU ITMO, Finland

Xin Wang, University of Southampton, U.K.

Stefan Wueller, RWTH Aachen, Germany

SELECTED PAPERS BOOK

A number of selected papers presented at WEBIST 2014 will be published by Springer-Verlag in a LNBIP Series book. This selection will be done by the Conference Chair and Program Chair, among the papers actually presented at the conference, based on a rigorous review by the WEBIST 2014 Program Committee members.

FOREWORD

This book contains the proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST 2014) which was organized and sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC). The conference was technically sponsored by the European Research Center for Information System (ERCIS).

The purpose of this Conference was to bring together researchers, engineers and practitioners interested in technological advances and business applications of web-based information systems. WEBIST had five main topic areas, covering different aspects of Web Information Systems, including "Internet Technology", "Web Interfaces and Applications", "Society, e-Business, e-Government", "Web Intelligence" and "Mobile Information Systems". We believe the proceedings here published demonstrate new and innovative solutions, and highlight technical problems in each field that are challenging and worthwhile.

The conference was also complemented with one special session, namely the Special Session on Business Apps - BA 2014 (chaired by Tim A. Majchrzak).

WEBIST 2014 received 153 paper submissions from 49 countries in all continents. A double-blind review process was enforced, with the help of 214 experts from the international program committee, all of them with a Ph.D. in one of the main conference topic areas. From these paper submissions only 23 papers were selected to be published and presented as full papers, i.e. completed work (12 pages in proceedings / 30' oral presentations) and 41 additional papers, describing work-in-progress as short papers for 20' oral presentation. Furthermore 35 short papers will be also presented as posters. The full-paper acceptance ratio was 15,03%, and the total oral paper acceptance ratio was 41,83%. These ratios denote a high level of quality, which we intend to maintain or reinforce in the next edition of this conference.

The high quality of the WEBIST 2014 programme is enhanced by four keynote lectures, delivered by experts in their fields, including (alphabetically): Andreas Pfeiffer (Hubject GmbH - joint venture of BMW Group, Bosch, Daimler, EnBW, RWE & Siemens, Germany), Fabien Gandon (INRIA, France), Steven Willmott (3scale, Spain) and Zakaria Maamar (Zayed University, United Arab Emirates).

Besides the proceedings edited by SCITEPRESS, a post-conference book will be compiled with extended versions of the conference´s best papers, and published by Springer-Verlag. Appropriate indexing has been arranged for the proceedings of WEBIST 2014 including Thomson Reuters Conference Proceedings Citation Index, INSPEC, DBLP, EI and Scopus. Furthermore, all presented papers will be available at the SCITEPRESS digital library.

The best contributions to the conference and the best student submissions were distinguished with awards based on the best combined marks of paper reviewing, as assessed by the

Program Committee, and the quality of the presentation, as assessed by session chairs at the conference venue.

In conjunction with a set of related conferences, namely SMARTGREENS and CLOSER, WEBIST had a new kind of session, named The European Project Space (EPS). The EPS aimed to provide insights into the research projects that are currently going on in Europe within the fields of web-based information systems, Cloud Computing and Smart Grids and Green IT Systems. This initiative intended to facilitate opportunities for knowledge and technology sharing, and establish the basis for future collaboration networks involving current project partners and interested conference delegates.

Building an interesting and successful program for the conference required the dedicated effort of many people. Firstly, we must thank the authors, whose research and development efforts are recorded here. Secondly, we thank the members of the program committee and additional reviewers for their diligence and expert reviewing. We also wish to include here a word of appreciation for the excellent organization provided by the conference secretariat, from INSTICC, which has smoothly and efficiently prepared the most appropriate environment for a productive meeting and scientific networking. Last but not least, we thank the invited speakers for their invaluable contribution and for taking the time to synthesize and deliver their talks.

Valérie Monfort
Université de Paris1 Panthéon Sorbonne, France

Karl-Heinz Krempels
RWTH Aachen University, Germany

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

Software is Eating the World, APIs are Eating Software IS-5
Steven Willmott

Typed Graphs and Linked Data - Modelling and Analyzing Social-semantic Web Data IS-7
Fabien Gandon

Accelerating the Electric Mobility Market Through New Services and an Efficient Market Mode IS-11
Andreas Pfeiffer

Enterprise 2.0 - Research Challenges and Opportunities IS-13
Zakaria Maamar

INTERNET TECHNOLOGY

FULL PAPERS

A Multitier Debugger for Web Applications 5
Manuel Serrano

Authentication and Authorisation for Widget-based Applications in a Loosely-coupled Web eLearning Environment 17
Jean-Noël Colin and Minh Tien Hoang

Annotated Trees and their Applications to XML Compression 27
Tomasz Müldner, Jan Krzysztof Miziołek and Tyler Corbin

Filter-enabled Binary XML Dissemination in Embedded Networks 40
Sebastian Käbisch and Richard Kuntschke

QoS-Aware Parameterized Semantic Matchmaking Framework for Web Service Composition 50
Salem Chakhar, Alessio Ishizaka and Ashraf Labib

Low Complexity Spatial Similarity Measure of GPS Trajectories 62
Radu Mariescu-Istodor, Andrei Tabarcea, Rahim Saeidi and Pasi Fränti

SocialSearch - A Social Platform for Web 2.0 Search 70
Claudio Biancalana, Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti

An Overview of Cloud Identity Management-Models 82
Bernd Zwattendorfer, Thomas Zefferer and Klaus Stranacher

Interoperability for Web Services based Smart Home Control Systems 93
Hannu Järvinen and Petri Vuorimaa

An Integrated Approach for Designing and Validating REST Web Service Compositions 104
Irum Rauf, Faezeh Siavashi, Dragos Truscan and Ivan Porres

SHORT PAPERS

On Metrics for Measuring Fragmentation of Federation over SPARQL Endpoints 119
Nur Aini Rakhmawati, Marcel Karnstedt, Michael Hausenblas and Stefan Decker

Delay-based Congestion Control Mechanism for Video Services - Mechanism including Backward Loading and Real-time Modes 127
Juha Vihervaara and Pekka Loula

An Approach to Transform Public Administration into SOA-based Organizations 135
J. Sedeño, C. J. Torrecilla-Salinas, M. J. Escalona and M. Mejías

A Flexible System for a Comprehensive Analysis of Bibliographical Data 143
Sahar Vahdati, Andreas Behrend, Gereon Schüller and Rainer Manthey

Implementing a Semantic Catalogue of Geospatial Data 152
Helbert Arenas, Benjamin Harbelot and Christophe Cruz

Testing Discovered Web Services Automatically 160
Pinar Karagoz and Selma Utku

Evaluating Web Service QoE by Learning Logic Networks 168
Natalia Kushik, Nina Yevtushenko, Ana Cavalli, Wissam Mallouli and Jeevan Pokhrel

From Image to Audio Watermarking Using Self-Inverting Permutations 177
Maria Chroni, Angelos Fylakis and Stavros D. Nikolopoulos

Implicit User Profiling in News Recommender Systems 185
Jon Atle Gulla, Arne Dag Fidjestøl, Xiaomeng Su and Humberto Castejon

Data Cube Computational Model with Hadoop MapReduce 193
Bo Wang, Hao Gui, Mark Roantree and Martin F. O'Connor

Distributed XML Processing over Multicore Servers 200
Yoshiyuki Uratani, Hiroshi Koide and Dirceu Cavendish

Are Model-driven Techniques Used as a Means to Migrate SOA Applications to Cloud Computing? 208
Miguel Botto, Javier González-Huerta and Emilio Insfran

Designing a General Architecture for Data Interchange 214
Alina Andreica, Josef Küng, Gabriela Şerban Czibulla and Christian Sacarea

Multimedia Retrieval based on Geometric Distance in Semi-structured Document 220
Sana Fakhfakh, Mohamed Tmar and Walid Mahdi

Extensible Data Management Architecture for Smart Campus Applications - A Crowdsourcing based Solution 226
Attila Adamkó and Lajos Kollár

Mobile Multi-modal Learning Facilitated through Twitter - A Case Study of Twitter Phenomenology with Graduate Level Nurses 233
Elizabeth M. La Rue, Lauren Terhorst and Kathleen G. McCafferty

FastTriaje: A Mobile System for Victim Classification in Emergency Situations 238
Alexandra Rivero-García, Candelaria Hernández-Goya, Iván Santos-González and Pino Caballero-Gil

SOCIETY, E-BUSINESS AND E-GOVERNMENT

FULL PAPERS

I Am off Then: Drivers of Travellers' Intentions to Book Trips Online - An Integrated Study on Technology Acceptance and Satisfaction 247
Maria Madlberger

TweetPos: A Tool to Study the Geographic Evolution of Twitter Topics 257
Maarten Wijnants, Adam Blazejczak, Peter Quax and Wim Lamotte

SHORT PAPERS

Experiences from and Attitudes towards Applying User Participation in Public e-Service Development 269
Jesper Holgersson and Eva Söderström

The Co-retweeted Network and Its Applications for Measuring the Perceived Political Polarization 276
Samantha Finn, Eni Mustafaraj and Panagiotis T. Metaxas

Analysis of LinkedIn Privacy Settings - Are they Sufficient, Insufficient or Just Unknown? 285
Pilar Manzanares-Lopez, Juan Pedro Muñoz-Gea and Josemaria Malgosa-sanahuja

Using PageRank for Detecting the Attraction between Participants and Topics in a Conversation 294
Costin Chiru, Traian Rebedea and Adriana Erbaru

An Implementation-independent Evaluation Model for Server-based Signature Solutions 302
Thomas Zefferer and Bernd Zwattendorfer

Business Process Improvement Methodology Based on Business Process Modelling - Case Study from Healthcare Sector 310
Faisal Aburub

A Model to Specify Inter-company Cooperation 318
Meriem Kermani, Mahmoud Boufaida and Emmanuel Paviot-Adet

The Effectiveness of Electronic Word of Mouth on Consumers' Perceptions of Adopting Products/Services - A Literature Review 324
Ali Reza Montazemi and Hamed Qahri Saremi

Communities of Practice Supporting Research and Technological Development - A Practical Case at Electrical Sector in Mexico 332
Yasmín Hernández, Salvador González-Castro and G. Arroyo-Figueroa

UnderstandIT: A Community of Practice of Teachers for VET Education 338
Maria De Marsico, Carla Limongelli, Filippo Sciarrone, Andrea Sterbini and Marco Temperini

Change Management in Service Oriented Virtual Organizations - A Structural Framework to Identify the Impacts and Triggers of Changes 346
Waeal J. Obidallah and Bijan Raahemi

An Investigation of Factors and Measurements for Successful e-Commerce Websites 354
Xiaosong Li, Wei Lian and Abdolhossein Sarrafzadeh

Who Runs the Internet? - Classifying Autonomous Systems into Industries 361
Annika Baumann and Benjamin Fabian

Exploring the Bitcoin Network 369
Annika Baumann, Benjamin Fabian and Matthias Lischke

Organizational Maturity and Projects Performance - The Mediation of Benefits Management 375
Jorge Gomes, Mário Romão, Helena Carvalho and Mário Caldeira

Technology-enabled Bullying & Adolescent Non-reporting - Breaking the Silence 381
Justin Connolly and Regina Connolly

Electronic Health Record Portal Adoption by Health Care Consumers - Proposal of a New Adoption Model 387
Jorge Tavares and Tiago Oliveira

Internet Science for Strategic Planning 394
Žiga Turk

Utilising Social Media Technology to Raise Brand Awareness in Higher Education 400
Baomin Qi and Lindsey Mackie

MOBILE INFORMATION SYSTEMS

FULL PAPERS

Cascading Information for Public Transport Assistance 411
Christian Samsel, Shirley Beul-Leusmann, Maximilian Wiederhold, Karl-Heinz Krempels, Martina Ziefle and Eva-Maria Jakobs

Evaluating Impact of Cross-platform Frameworks in Energy Consumption of Mobile Applications 423
Matteo Ciman and Ombretta Gaggi

SHORT PAPERS

Robots Collaboration based on Cloud Robotics System for Daily - Emergency Life 435
Buribayeva Gulban and Taizo Miyachi

Provision of Personalized Data via Mobile Web Services in eHealth Scenarios 441
Marc Jansen, Abbas Siddiqui and Oliver Koch

AUTHOR INDEX 447

INVITED SPEAKERS

KEYNOTE SPEAKERS

Software is Eating the World, APIs are Eating Software

Steven Willmott
3scale, Spain

Abstract: Software is becoming increasingly critical in many industries - from optimizing processes and information sharing, to powerful data analysis and embedded controllers. At the same time however, this software is becoming ever more componentzied and inter-connected. Using core web technologies such as HTTP, XML, JSON and others, Web APIs are becoming the glue that weaves together these multiplying software components. This talk covers how software can often dramatic and change the fortunes of companies in a sector as well as how APIs and related Web Information technologies are part of this big transition.

BRIEF BIOGRAPHY

Steven was previously the research director of one of the leading European research groups in Europe on distributed systems and Artificial Intelligence at the Universitat Politècnica de Catalunya in Barcelona, Spain. He brings 10 years of technical experience in Web Services, Semantic Web, network technology and the management of large-scale international R&D teams.

Typed Graphs and Linked Data
Modelling and Analyzing Social-semantic Web Data

Fabien Gandon
INRIA, Univ. Nice Sophia Antipolis, CNRS, I3s, UMR 7271, 06900 Sophia Antipolis, France
Fabien.Gandon@Inria.Fr

Keywords: Semantic Web, Social Web, Typed Graphs, Knowledge Representation, Ontologies.

Abstract: This extended abstract summarizes the subject of an invited talk at WebIST 2014 that introduces methods, models and algorithms studied by the Wimmics research lab to bridge formal semantics and social semantics on the Web.

1 HYBRID WEB COMMUNITIES

The Web is now a virtual place where users and software interact in hybrid communities. These large scale interactions create many problems in particular the one of reconciling formal semantics of computer science (e.g. logics, ontologies, typing systems, etc.) on which the Web architecture is built, with soft semantics of people (e.g. posts, tags, status, etc.) on which the Web content is built.

The Wimmics research lab studies methods, models and algorithms to bridge formal semantics and social semantics on the Web. The name of the lab stands for "web-instrumented man-machine interactions, communities, and semantics" and we address this problem focusing on the characterization of typed graphs formalisms to model and capture these different pieces of knowledge and hybrid operators to process them jointly.

The approaches we introduce rely on graph-oriented knowledge representation, reasoning and operationalization to model and support actors, actions and interactions in web-based epistemic communities. The research results are applied to support and foster interactions in online communities and manage their resources.

This extended abstract introduces: the need to combine different kinds of graph-based representations (section 2), the importance of having a multidisciplinary approach (section 3); and the interest of relying on typed-graph knowledge representations. The last section provides a number of references.

2 FORMAL-SOCIAL SEMANTICS

The initial graph of linked pages of the Web has been joined by a growing number of other graphs: social networks, workflows, navigation trails, automatas of distributed services, linked open data clouds, etc. And these graphs interact in all sorts of ways influencing the life-cycle of each other.

Not only do we need means to represent and analyse each kind of graphs, we also need the means to combine them and to perform multi-criteria analysis on their combination.

We propose to address this problem by focusing on the characterization of (a) typed graphs formalisms to model and capture these different pieces of knowledge with their links and (b) hybrid operators to process them jointly. We especially consider the problems that occur in such structures when we blend formal stable semantic models and socially emergent and evolving semantics.

3 MULTIDISCIPLINARY MODEL

Our research follows two main directions combining two complementary types of contributions:

- First research direction: to propose multidisciplinary approach to analyse and model the many aspects of these intertwined information systems, their communities of users and their interactions;
- Second research direction: to propose formalizations of the previous models and reasoning algorithms on these models

providing new analysis tools and indicators, and supporting new functionalities and better management.

In a nutshell, the first research direction looks at models of systems, users, communities and interactions while the second research direction considers formalisms and algorithms to represent them and reason on their representations.

In the short term we intend to survey, extend, formalize and provide reasoning means over models representing systems, resources, users and social links in the context of social semantic web applications.

In the longer term we intend to extend these models (e.g. dynamic aspects), unify their formalisms (dynamic typed graphs) and propose mixed operations (e.g. metrical and logical reasoning) and algorithms to scale them (e.g. random walks) to support the analysis of epistemic communities' structures, resources and dynamics.

Ultimately our goal is to provide better collective applications on the web of data and the semantic web with two sides to the problem: (1) improve access and use of the linked data for epistemic communities and at the same time (2) use typed graph formalisms to represent the web resources, users and communities and reason on them to support their management.

4 TYPED GRAPHS

The work presented in this invited talk reports on models and algorithms to bridge formal semantics and social semantics by formalizing and reasoning heterogeneous semantic graphs.

The models we design include: users, their profiles, their requirements, their activities and their contexts; social links, social structures, social exchanges and processes; conceptual models including ontologies, thesauri, and folksonomies. Whenever possible these models are formalized and published according to standardized web formalisms and may motivate research and suggestions on extending these standards. The main goal here is to improve the understanding the systems have of the communities, resources and activities of their users. We focus on formalizing the models as unified typed graphs in order for software to exploit them jointly in their processing while still being able to use the specificities of each one of them. The schemas and datasets are published as linked data following the web architecture principles and W3C standards.

The algorithms we designed include: typed graphs indexing, reasoning and searching; hybrid processing merging logical inferences, rules and metrical inferences; approximation and propagation algorithms; distributed querying and reasoning. We propose algorithms (in particular graph-based reasoning) and approaches (in particular knowledge-based approaches) to process the mixed representations we obtain. We are especially interested in allowing cross-enrichment between them and in exploiting the life cycle and specificities of each one to foster the life-cycles of the others. These algorithms are published, implemented and distributed as part of a generic open source platform and library called Corese/KGram.

REFERENCES

Wimmics web site, 2014, http://wimmics.inria.fr/

Marie, N., Ribiere, M., Gandon, F., Rodio, F., 2013. *Discovery Hub: on-the-fly linked data exploratory search*, I-Semantics.

Villata, S., Costabello, L., Gandon, F., Faron-Zucker, C., Buffa, M., 2013, *Social Semantic Network-based Access Control*, Security and Privacy Preserving in Social Networks, Lecture Notes in Social Networks, Springer

Cojan, J., Cabrio, E., Gandon, F., 2013. *Filling the Gaps Among DBpedia Multilingual Chapters for Question Answering*, ACM Web Science Conference

Buffa, M., Ereteo, G., Limpens, F., Gandon, F., 2013. *Folksonomies and Social Network Analysis in a social Semantic Web*, 39th International Conf. on Current Trends in Theory and Practice of Computer Science.

Corby, O., Gaignard, A., Faron-Zucker, C., Montagnat, J., 2012. *KGRAM Versatile Data Graphs Querying and Inference Engine*, Proc. IEEE/WIC/ACM International Conference on Web Intelligence

Corby, O., Faron-Zucker, C., 2010. *The KGRAM Abstract Machine for Knowledge Graph Querying*, IEEE/WIC/ACM International Conference.

Erétéo, G., Buffa, M., Gandon, F., Corby, O., 2009. *Analysis of a Real Online Social Network using Semantic Web Frameworks.* International Semantic Web Conference, ISWC'09.

Monnin, A., Limpens, F., Gandon, F., Laniado, D., 2010. *Speech acts meets tagging: NiceTag ontology*, AIS SigPrag International Pragmatic Web Conference Track, I-Semantics.

Basse, A., Gandon, F., Mirbel, I., Lo, M., 2010. *Frequent Graph Pattern to Advertise the Content of RDF Triple Stores on the Web*, Web Science Conference

BRIEF BIOGRAPHY

Fabien Gandon is Senior Research Scientist and HDR in Informatics and Computer Science at INRIA and he is the Leader of the Wimmics team at the Sophia-Antipolis Research Center. He is also a member of the World-Wide Web Consortium (W3C) where he participates in several standardization groups. His professional interests include: Web, Semantic Web, Social Web, Ontologies, Knowledge Engineering and Modelling, Mobility, Privacy, Context-Awareness, Semantic Social Network/ Semantic Analysis of Social Network, Intraweb. He previously worked for the Mobile Commerce Laboratory of Carnegie Mellon in Pittsburgh.

Accelerating the Electric Mobility Market Through New Services and an Efficient Market Mode

Andreas Pfeiffer
Hubject GmbH - joint venture of BMW Group, Bosch, Daimler, EnBW, RWE & Siemens, Germany

Abstract: How to connect different providers of electric mobility services and ensure an efficient design? An eRoaming platform as well as the contractual and technical framework required to implement this concept throughout Europe have already been developed by the Berlin-based joint venture Hubject. In addition an open interface protocol, the "Open InterCharge Protocol" (OICP), has been defined and can be accessed online by market participants for free since March 2013. Hubject GmbH is a joint venture formed by the BMW Group, Bosch, Daimler, EnBW, RWE and Siemens based in Berlin. The company was formed in 2012 and operates a cross-industry business and IT platform connecting infrastructure, service and mobility providers throughout Europe.

BRIEF BIOGRAPHY

Andreas Pfeiffer studied Business Administration at the RWTH Aachen University after training to become an information technology consultant. Upon successful completion of the degree, he worked as an organisational developer for an IT service provider. In 2007, he moved to Energieversorgungs- und Verkehrsgesellschaft mbH Aachen (E.V.A.) where he worked as a corporate developer. He managed the topic "electric mobility" for the Energieversorgungs- und Verkehrsgesellschaft mbH Aachen (E.V.A.) from 2008 to 2011 and was Managing Director of smartlab Innovationsgesellschaft mbH in Aachen from 2010 to 2012. Andreas Pfeiffer is Managing Director of Hubject GmbH, a joint venture of BMW Group, Bosch, Daimler, EnBW, RWE and Siemens based in Berlin, since August 2012.

Enterprise 2.0
Research Challenges and Opportunities

Zakaria Maamar
Zayed University, Dubai, United Arab Emirates

EXTENDED ABSTRACT

Today's enterprises are caught in the middle of a major financial crisis that is undermining their profit and growth and even jeopardizing their survival. To respond to this crisis, enterprises have launched different initiatives to improve their business processes and align their strategies with market needs, for example. Regular enterprise applications implement structured business processes in which the steps to perform are well defined. While this is beneficial for enterprises, there is little room for creativity in performing such processes without triggering a complex re-thinking process. Usually this process takes time to implement and analyze its effects, which is sometimes late and inefficient due to business constant changes. Several decisions are to be made on the fly and could tap into the large volume of data that Web 2.0 (social) applications (e.g., social networks, blogs, and wikis) generate.

Social applications rely on users' ability and willingness to interact, share, and recommend. However the richness and complexity of information in these applications pose challenges for enterprises on how to capture and structure this information for future use while preserving user privacy and information sensitivity. The social "fever" has caught every single activity of people's daily life ranging from sharing live experiences online to seeking feedback on any matter like what to wear for a special occasion. A report encourages enterprises to allow their employees to embrace social applications in order to establish and foster contacts with their colleagues, customers, and suppliers (B. Peter and R. Reeves. Network Citizens: Power and Responsibility at Work, Demos, 29 October 2008, www.demos.co.uk/publications/networkcitizens). This should impact in a positive way productivity, business development, and collegiality as long as these applications are used in a controlled way.

This presentation offers a research roadmap on the challenges and opportunities that Enterprise 2.0 (aka Social Enterprise) has to deal with and tap into, respectively. Contrary to traditional enterprises with a top-down command flow and bottom-up feedback flow, these flows in Enterprise 2.0 cross all levels and in all directions bringing all people together for the development of creative and innovative products and services. McAfee was the first to introduce the term Enterprise 2.0 as the use of emergent social software platforms within or between companies and their partners or customers (A. P. McAfee, Enterprise 2.0: The Dawn of Emergent Collaboration, MIT Sloan Management Review, 47(3), 2006). Enterprise social software should work hand-in-hand with regular business processes to ensure Enterprise 2.0 success (BLUEKIWI, The State of Social Business 2013: The Maturing of Social Media into Social, Business, blueKiwi report, 2013): "Enterprise 2.0 only works if it is part of a business process. It's great to work in new ways, but it's not enough. To make it real, it has to be very practical". Enterprise 2.0 should develop techniques and guidelines that would allow weaving social relationships (e.g., collegiality, fairness, and trustworthiness) into their operation. This should lead into a new form of business processes in the enterprise that reinforce the fact that employees establish and maintain social networks of contacts, rely on some privileged contacts when needed, and form with other peers strong and long lasting social collaborative groups. In today's economies, an enterprise's ability to sustain its growth and competitiveness depends on how well it manages from a social perspective its communications with stakeholders that are customers, suppliers, competitors, and partners (Y. Badr and Z. Maamar. Can Enterprises Capitalize on Their Social Networks? Cutter IT Journal, Special Issue on Measuring the Success of Social Networks in the Enterprise, 22(10), October 2009).

BRIEF BIOGRAPHY

Zakaria Maamar is a Professor and Assistant Dean in the College of Information Technology at Zayed University in Dubai, United Arab Emirates. His research interests are primarily related to service-oriented computing, social computing, and system interoperability. Dr. Maamar graduated for his M.Sc. and Ph.D. in Computer Sciences from Laval University in Canada in 1995 and 1998, respectively. Related Experiences: Organizer of several workshops in the past Quan Z. Sheng, Aikaterini Mitrokotsa, Sherali Zeadally, Zakaria Maamar (Eds.): RFID Technology-Concepts, Applications, Challenges, Proceedings of the 4th International Workshop, IWRT 2010, In conjunction with ICEIS 2010, Funchal, Madeira, Portugal, June 2010. SciTePress 2010, ISBN 978-989-8425-11-9 Soraya Kouadri Mostéfaoui, Zakaria Maamar, George M. Giaglis (Eds.): Ubiquitous Computing, Proceedings of the 3rd International Workshop on Ubiquitous Computing, IWUC 2006, In conjunction with ICEIS 2006, Paphos, Cyprus, May 2006. INSTICC Press 2006, ISBN 978-972-8865-51-1

INTERNET TECHNOLOGY

FULL PAPERS

A Multitier Debugger for Web Applications

Manuel Serrano
Inria Sophia Méditerranée, 2004 route des Lucioles, F-06902 Sophia Antipolis, France
Manuel.Serrano@inria.fr

Keywords: Web Debugging, Web Programming, Functional Programming.

Abstract: Debugging Web applications is difficult because of their distributed nature but also because of the programming languages and tools commonly used to develop them. Taking benefit of the multitier aspect of the Hop programming language, we have built a new debugger for Web applications that copes with the server-side and the client-side of the executions. Its advantage over most debuggers for the Web is that it reports the full stack trace containing all the server-side and client-side frames that have conducted to an error. An error is reported on its actual position on the source code, wherever it occurs on the server or on the client.
To help detecting errors as early as possible, the Hop debugger is accompanied with a debugging execution mode where types are checked before data structures are accessed, argument numbers are verified before functions are called, and array bounds are checked before vectors are accessed. Combining the debugger and the debugging mode makes errors of Web applications easier to understand and easier to localize. Hopefully they also become easier to fix.

1 INTRODUCTION

The distributed nature of Web applications makes debugging difficult. The programming languages and tools commonly used make it even more complex. Generally the server-side and the client-side are implemented in different settings and the debugging is treated as two separated tasks: on the one hand, the debugging of the server, on the other hand, the debugging of the client. Most studies and tools focus on this latter aspect. They concentrate on the debugging of JavaScript in the browser. Although useful, this only addresses one half of the problem. Considering the debugging of Web applications as a whole raises the following difficulties:

- As the server-side and the client-side are generally implemented in different languages, debuggers for the Web do not capture the whole execution of the application. Programming the server and the client in the same language helps but is not sufficient to let the debugger expose a coherent view of the whole execution as this also demands a runtime environment that enforces consistent representations of data structures and execution traces.
- The JavaScript tolerant semantics tends to defer errors raising. For instance, calling a function with an insufficient number of arguments may lead to filling a data structure with the unexpected `undefined` value which, in turn, may raise a type error when accessed. The *distance* between the error and its actual cause may be arbitrarily long, which can make the relation between the two difficult to establish.
- The JavaScript event loop used for the GUI splits the execution into unrelated callback procedures which get called upon event receipts. When an error occurs, the active stack trace only contains elements relative to the current callback invocation. It is oblivious of the context of the callback. Understanding the cause of the error is then uneasy.

Pursuing our research on multitier programming for the Web, we have built a programming environment that eliminates most of these problems.

- When an error is raised, the full stack trace is reported. This stack trace may contain server stack frames, client stack frames, or both. We call this a *multitier stack trace*.
- When an error occurs, either on the client or on the server, its source location is reported by the debugger.
- In *debugging mode*, types, arities, and array bounds, are strictly enforced on the server and on the client. Hence, when the execution of the pro-

gram deviates from the formal semantics of the language, an error is raised immediately.

This paper presents this debugger and exposes the salient aspects of its implementation.

1.1 Debugging Web Applications

Most studies of the debugging of Web applications consider the client-side of the applications. Beside an early debugger for CGI applications by Vckovski (Vckovski, 1998), most efforts concentrate on creating JavaScript debuggers for the Web browsers. Feldman and Sharma have both developed a remote debugger for WebKit (Feldman, 2011; Sharma, 2012). Mickens has developed a browser-agnostic JavaScript debugger (Mickens, 2012). These tools offer facilities for inspecting the JavaScript stack frame and for breakpointing. They do not address the problem of debugging Web applications globally. These studies are complementary with our effort as Hop can use them to step inside the client side execution, provided they support the *source map* specification (Lenz and Fitzgerald, 2011).

A previous study by E. Schrock (Schrock, 2009) has identified many difficulties posed by JavaScript when debugging Web applications. *"In the early days, the [JavaScript]'s ability to handle failure silently was seen as a benefit. If an image rollover failed, it was better to preserve a seamless Web experience than to present the user with unsightly error dialogs.*

This tolerance of failure has become a central design principle of modern browsers, where errors are silently logged to a hidden error console... Now, at best, script execution failure results in an awkward experience. At worst, the application ceases to work or corrupts server-side state. Tacitly accepting script errors is no longer appropriate, nor is a one-line number and message sufficient to identify a failure in a complex AJAX application. Accordingly, the lack of robust error messages and native stack traces has become one of the major difficulties with AJAX development today... Ideally, we would like to be able to obtain a complete dump of the JavaScript execution context[.]". This is exactly what Hop brings. In addition to offering multitier execution stacks that reflect the states of the server and the client, the Hop-to-JavaScript compiler (Loitsch and Serrano, 2007) generates codes which enable early error detection. It also generates `ECMA-262-5` *strict mode* (ECMA, 2009) code which helps detecting undeclared variables and functions that can be located inside JavaScript code linked against Hop applications.

1.2 The Hop Programming Language

Hop has been presented in several publications (Serrano et al., 2006; Boudol et al., 2012; Serrano and Berry, 2012). We only remind its essential aspects and show some examples that should be sufficient to understand the rest of the paper.

Hop is a Scheme-based multitier functional language. The application server-side and client-side are both implemented within a single Hop program. Client code is distinguished from server code by prefixing it with the syntactic annotation '`~`'. Server-side values can be injected inside a client-side expression using a second syntactic annotation: the '`$`' mark. On the server, the client-side code is extracted, compiled on-the-fly into standard JavaScript, and shipped to the client. This enables Hop clients to be executed by unmodified Web browsers.

Except for multitier programming, the standard Web programming model is used by Hop. A server-side Hop program builds an HTML tree that creates the GUI and embeds client-side code into scripts, then ships it to the client. AJAX-like service-based programming is made available by service definitions, a service being a server-side function associated with a URL. The `with-hop` form triggers execution of a service. Communication between clients and servers is automatically performed by the Hop runtime system, with no additional user code needed.

The Hop Web application `fib-html` below consists of a server-built Web page displaying a three-rows table whose cells enumerate positive integers. When a cell is clicked, the corresponding Fibonacci value is computed on the client and displayed in a popup window. Note the '`~`' signs used in lines *3*, *8*, *9*, and *10* which mark client-side expressions.

```
1: (define-service (fib-html)
2:  (<HTML>
3:   ~(define (fib x) ;; client code
4:      (if (< x 2)
5:          1
6:          (+ (fib (- x 1)) (fib (- x 2)))))
7:   (<TABLE>
8:    (<TR> (<TD> "fib(1)" :onclick ~(alert (fib 1))))
9:    (<TR> (<TD> "fib(2)" :onclick ~(alert (fib 2))))
10:   (<TR> (<TD> "fib(3)" :onclick ~(alert (fib 3)))))))
11:
```

Let us modify the example to illustrate some Hop niceties. Instead of building the rows by hand, we let Hop compute them. The new Hop program uses the `(iota 3)` expression (line *9*) that evaluates to the list (1, 2, 3) and the `map` functional operator that applies a function to all the elements of a list. The `$i` expression (line *8*) denotes the value of `i` on the server at HTML document elaboration time.

```
1: (define-service (fib-html)
2:    (<HTML>
3:       ~(define (fib x) ...)
4:       (<TABLE>
5:          (map (lambda (i)
6:                  (<TR>
7:                     (<TD> "fib(" i ")"
8:                        :onclick ~(alert (fib $i)))))
9:               (iota 3)))))
```

Before delivery to a client, the server-side document is compiled on the server into regular HTML and JavaScript. This produces the following document [1]:

```
 1: <!DOCTYPE HTML>
 2: <html>
 3:   <head> <meta ...>
 4:     <script src='/usr/local/share/hop/hop_s.js'>
 5:   </head>
 6:   <script>
 7:     function fib(x) {if(x<2) return 1; else ...}}
 8:   </script>
 9:   <table>
10:     <tr><td onclick="alert(fib(1))">fib(1)</td></tr>
11:     <tr><td onclick="alert(fib(2))">fib(2)</td></tr>
12:     <tr><td onclick="alert(fib(3))">fib(3)</td></tr>
13:   </table>
14: </html>
```

This program can then be executed by all standard browsers.

If for some reason, the programmer wants the `fib` calls to be evaluated on the server, three modifications are needed: *i)* a service must be defined as the client cannot access directly server-side functions, *ii)* the definition of the `fib` function must be migrated to the server, and *iii)* a `with-hop` remote call must be introduced:

```
 1: (define-service (fib-html)
 2:   (<HTML>
 3:     (<TABLE>
 4:       (map (lambda (i)
 5:             (<TR>
 6:                (<TD> "fib(" i ")"
 7:                   :onclick ~(with-hop ($fib-svc $i)
 8:                               alert))))
 9:        (iota 3)))))
10:
11: (define-service (fib-svc n)
12:    (define (fib x) ...)
13:    (fib n))
```

These examples illustrate the flavor of Web programming with Hop. Many more are available on the Hop web site[2].

In *debugging* mode, Hop generates JavaScript code that enforces types consistency, arity correctness, and array bounds. In this mode, all functions must be called with a number of arguments compatible with their declaration, all accesses to the data structures and to the arrays are verified by the runtime system. By experience, we have found strict checking an effective method for correcting errors that are otherwise difficult to understand. Since preventing runtime errors has a cost, Hop also supports a *production* mode where types, arities, and bounds are not enforced at runtime. This yields faster but *unsafe* executions. In production mode, executions are not guaranteed to comply with the Hop's formal semantics (Boudol et al., 2012).

[1] Some generated programs have been manually modified to fit the paper layout constraints.

[2] `http://hop.inria.fr`

1.3 Organization of the Paper

The paper is organized as follows. Section 2 presents some debugging scenarios. Section 3 sketches the implementation of the debugger and the debugging mode. A comparison of the performance between the *debugging* and *production* modes is then presented. Section 4 presents the related work.

2 DEBUGGING SCENARIOS

In this section we show the error reports produced by the Hop *debugging* mode in typical erroneous situations. The reports are accessible on the client, *i.e.*, the Web browser, and on the server, which prints them on its console.

2.1 Client-side Error

Let us first consider a type error that involves client-side and server-side computations. The server-side program elaborates a single-button Web page which invokes the client-side function `my-callback` when clicked (line *10*).

```
 1: (module bug1_server
 2:    ~(import bug1_client)
 3:    (export bug1-svc))
 4:
 5: (define-service (bug1)
 6:    (<HTML>
 7:       (<HEAD>
 8:          :jscript (service-resource bug1 "bug1.scm"))
 9:       (<BUTTON> :id "my-button"
10:          :onclick ~(my-callback)
11:          "click me to raise the error")))
12:
13: (define-service (bug1-svc a b)
14:    (vector a b))
```

The function `my-callback` is implemented in a separate client-side module.

```
 1:(module bug1_client
 2:   $(import bug1_server)
 3:   (export (my-callback)))
 4:
 5:(define (my-callback)
 6:   (with-hop ($bug1-svc 11 12)
 7:      my-type-error))
 8:
 9:(define (my-type-error l)
10:   (car l))
```

The client-side function `my-callback` (line *5*) calls the server service `bug1-svc`. When the service completes, the execution resumes on the client by invoking the function `my-type-error`. This function is called with the vector its receives from the service. This is wrong because its wants a pair. The vector/list type mismatch is reported as:

```
File "bug1.scm", line 10, character 3:
#   (car l))
#   ^
*** CLIENT ERROR: http://localhost:8888/hop/bug1, car:
Type "pair" expected, "vector" provided -- #(11 12)
   1. ~my-type-error@bug1_client, bug1.scm:9
With-Hop trace:
   2. ~(with-hop (bug1-svc...) ...), bug1.scm:6
   3. ~my-callback@bug1_client, bug1.scm:5
   4. ~button#my-button.onclick, bug1.hop:10
```

The error report shows the position in the source file of the error and the complete stack trace. Client-side frames are prefixed with the ~ sign. (Here, all stack frames are client frames.) When the type error is raised in the client-side function `my-type-error` the active stack trace only contains one frame denoting the invocation of the function `my-type-error`. However, the report also shows the context from which `my-type-error` has been called. It shows that a click on the button defined line *9* of `bug1_server` module has called the client-side function `my-callback` which, in turn, has invoked the remote service `bug1-svc`.

The call trace makes explicit the whole execution flow which has conducted to the error on the client. As it also makes explicit the network traversal (frame #2) that took place before `my-type-error` is called, it is easier to understand which actual computation conducted to the type error.

This scenario also illustrates an important difference between Hop and JavaScript. In JavaScript a function equivalent to `my-type-error` would silently return the `undefined` value. In Hop *debugging* mode, the type error is signaled as soon as the illegal access gets executed.

2.2 Server-side Error

Let us modify the definition of the previous service to introduce a server-side error:

```
13:(define-service (bug2-svc a b)
14:   (vector (car a) b))
```

This is wrong as the service `bug2-svg` is passed an integer for the argument `a` while it expects a pair. The new error report is as follows:

```
File "bug2.hop", line 14, character 301:
#   (vector (car a) b))
#           ^
*** SERVER ERROR:car
Type "pair" expected, "int" provided -- 11
    1. \@bug2-svc, bug2.hop:13
    2. &pool-scheduler1965,
With-Hop trace:
    3. ~(with-hop (bug2-svc...) ...), bug2.scm:6
    4. ~my-callback@bug2_client, bug2.scm:5
    5. ~button#my-button.onclick, bug2.hop:10
```

The report locates the error inside the `bug2-svc` service (services are prefixed with `\@` to distinguish them from regular functions). The stack trace shows server-side stack frames and the client-side context that has yielded the service invocation. This time again the error is easy to follow and to understand as the complete trace before the error is exposed. The computation started with a user click on the client. The click action has been followed by a service invocation, which has raised the server-side error.

2.3 Putting Together

The Hop debugger keeps track of all the callbacks of the client-side program. Let the callback be associated with a service invocation as seen before, with a GUI event (mouse move, key press, ...), with a server side event (a high level facility supported by Hop built on top of websockets), or with a timer, Hop generates a dedicated entry in the stack trace. Callback traces can be combined without restriction. For instance, suppose a service call wrapped in an `after` expression (a mere JavaScript `setInterval` wrapper) as follows:

```
5:(define (my-callback)
6:   (after 1000
7:      (lambda ()
8:         (with-hop ($bug2-svc 11 12)
9:            my-type-error))))
```

The new error trace shows a *with-hop* trace preceded by an *after* trace:

```
File "bug2.hop", line 14, character 295:
#   (vector (car a) b))
#           ^
*** SERVER ERROR:car
Type "pair" expected, "bint" provided -- 11
   1. \@bug2-svc, bug2.hop:13
   2. &pool-scheduler1926,
With-Hop trace:
   3. ~(with-hop (bug2-svc...) ...), bug3.scm:8
   4. ~lambda, bug3.scm:7
After trace:
   5. ~after, /usr/local/share/hop/hop_s.js:8254
   6. ~my-callback@bug3_client, bug3.scm:5
   7. ~button#my-button.onclick, bug2.hop:10
```

A Hop stack trace may also contain JavaScript frames as JavaScript functions are treated as Hop functions. This can be observed in the previous stack trace with the `after` function (frame #5) which is a plain JavaScript function implemented in the Hop standard client library (the file `hop_s.js`).

Beside full stack traces, the Hop debugger also supports stepping. In the current version, the server-side stepper and client-side stepper work separately and only the client-side stepper is fully operational. The server-side is in progress, being developed using the techniques presented in (Kellomäki, 1993). The implementation of the client-side stepper is briefly discussed in Section 3.1.

3 IMPLEMENTATION

The Hop debugger relies on three elements: *i)* stepper, *ii)* construction the stack traces, and *iii) debugging mode* where types, array bounds, and function arities are verified at runtime. These are presented in this section.

3.1 Implementing the Stepper

The server-side stepper is currently being developed using the techniques established for Lisp-like languages (Kellomäki, 1993). The client-side stepper reuses the native stepper available in modern browsers. The Hop-to-JavaScript compiler produces *Source Map* tables (Lenz and Fitzgerald, 2011) that let browsers step in the client side code using the original Hop source code instead of the generated JavaScript one.

3.2 Constructing Stack Traces

Hop client-side programs are compiled into *natural* JavaScript programs, namely, Hop functions are mapped into JavaScript functions, and Hop variables are mapped into JavaScript variables. Hence, obtaining the client-side part of the Hop stack frames is similar to obtaining a plain JavaScript stack frames, whose technique is well known (Schrock, 2009; Mickens, 2012; Sharma, 2012). It relies on two observations: first, JavaScript exception objects contain stack information; second, there are four different contexts in which codes get executed:

1. the global context while loading the page;
2. event listeners (GUI or server events);
3. timeouts and intervals;
4. remote service callbacks (XmlHTTPRequest).

To obtain a stack trace, the runtime environment installs exception handlers on these four contexts, it intercepts exceptions, and extracts their stack information. Older techniques based on the two special JavaScript variables `caller` and `callee` are now impractical as JavaScript *strict mode* used by Hop does not support these variables.

3.2.1 Constructing Multitier Stack Traces

The *multitier stack trace* describes the current computation and the context in which it has been initiated. Contexts are computed as follows:

1. The context of a global top-level JavaScript evaluation is empty.
2. The context of a DOM listener specified as an attribute of an HTML node consists of a description of the node and a description of the event the listener is attached to.
3. The context of an event listener dynamically attached to a DOM event consists of a description of the DOM node plus the stack trace active at the moment the listener is attached.
4. The context of a timeout consists of the concatenation of the context and the stack trace active when the callback is registered, and a description of the timeout itself.
5. The context of a service call (`with-hop`) consists of the concatenation of the context and the stack trace active when the remote call is spawn and a description of the called service.

Hop stores the active context in the JavaScript global variable `hop_current_stack_context`, which is updated each time a new callback is registered. This approach is correct because JavaScript execution is single-threaded and because callbacks always run up to completion (*i.e.*, they are never preempted).

Let us illustrate the construction of the multitier stack trace on two actual examples. First, let us show

the compilation of an HTML button declaration as found in the examples of section 2.1.

```
(<BUTTON> :id "my-button" :onclick ~(my-callback)
   "click me to raise the error")
```

The *production mode* compilation merely consists in mangling the Hop identifier to map it into the JavaScript identifiers space:

```
<button id='my-button'
  onclick='BGl_myzd2callbackzd2zzbug3_clientz00()'>
  click me to raise the error
</button>
```

The possibility to change the compilation schema according to external configuration is a benefit of the Hop approach where the JavaScript code is generated on demand by an on-the-fly compiler. Switching from production mode to debugging mode and *vice versa* merely requires switching on and off a Hop compiler flag. In *debugging mode*, the compilation of the button is changed for:

```
1: <button id='my-button'
2:   onclick='hop_callback(
3:     function () {BGl_myzd2callbackzd2zzbug3_clientz00()},
4:     hop_callback_html_context_handler(
5:       "button#my-button.onclick", "bug2.hop", 205))
6:   .call(this)'>
7:    click me to raise the error
8: </button>
```

This code constructs an HTML context that stores the source location of the button (line *4*). It calls the library the library function `hop_callback` (line *2*). The wrapped callback returned from `hop_callback` is then called with the `this` parameter of the HTML statement (line *6*).

The `hop_callback` function sets the global exception context (line *2*) and it wraps the user callback (`proc`) into a context aware callback (line *3*). This wrapped callback installs an error handler (line *4*) which signals potential errors in the context that was active when the callback has been installed (the variable `ctx` line *7*).

```
 1: function hop_callback( proc, ctx ) {
 2:    hop_current_stack_context = ctx;
 3:    return function() {
 4:       try {
 5:          return proc.apply( this, arguments );
 6:       } catch( e ) {
 7:          hop_callback_handler( e, ctx );
 8:       }
 9:    }
10: }
```

The function `hop_callback_handler`, which is invoked when an error is raised at execution time, simply extracts the stack trace it finds in the exception and the stack trace is builds from the callback context:

```
1: function hop_callback_handler( e, ctx ) {
2:    var estk = hop_get_exception_stack( e );
3:    var cstk = hop_get_context_stack( ctx );
4:    report_stack_frame( estck );
5:    report_stack_frame( cstck );
6: }
```

To show how contexts are accumulated, let us study the implementation of the Hop `after` function. It works similarly to `with-hop` but it is simpler to understand as `with-hop` carries its own complexity independently of stack contexts. The base implementation of `after` is as follows:

```
function after( timeout, proc ) {
   if( hop_debug() >= 0 ) { /* debugging code */ }
   var i = setInterval(
      function() { clearInterval( i ); proc() },
      timeout );
}
```

When debugging is enabled (`hop_debug() >= 0` is true), the extra following code is executed:

```
 1: /* debugging code */
 2: if( !sc_isNumber( timeout ) )
 3:    sc_typeError( "after", "integer", timeout, 1 );
 4: if( !("apply" in proc) )
 5:    sc_typeError( "after", "procedure", proc, 1 );
 6: try {
 7:    throw new Error( "after" );
 8: } catch( e ) {
 9:    var ctx = sc_cons( "After trace:",
10:       hop_append_context(
11:          hop_get_exception_stack( e ),
12:          hop_current_stack_context ) );
13:    proc = hop_callback(hop_arity_check(proc, 0), ctx);
14: }
```

First, initial type tests (lines *2* and *4*) are executed. Then, before calling the `setInterval` JavaScript function, an exception is raised to capture the current execution trace (the function `hop_get_exception_stack`.) This stack trace is concatenated to the context active when the function `after` is called. Line *13*, the call to the function `hop_arity_check` checks if the callback provided on the call site has a correct arity. If not, it throws an exception.

The proposed implementation of `after` breaks tail recursions. Programming patterns as:

```
(let loop ()
   (after delay
      (lambda () ... (loop))))
```

blow the memory because the contextual stack is augmented each time a new iteration of the loop is executed. Several *ad-hoc* solutions are possible to workaround this problem. The one implemented in Hop consists in checking the top of the contextual stack. If it is already an *after* frame, nothing is pushed on the stack. Otherwise a new frame is pushed as already described. Although simplistic, we have found

this solution sufficient and convenient to debug tail-recursive programs. If needed in the future, smarter solutions will be envisioned.

Handling the context stack on the server is simpler. When a client invokes a service, it serializes the context stack and ships it along the service arguments. The server protects the execution of its service with a handler that appends the execution trace of the exception to the client-side context. This augmented context is returned to the client if a server-side error occurs.

3.2.2 Pretty-printing the Stack Trace

Pretty printing stack traces requires the debugger to identify correctly Hop stack frames and to map the actual locations of the generated JavaScript file into the user source codes. In this process, a stack frame such as:

```
at BGl_myzd2callbackzd2zzbug1_clientz00 (bug1.js:4:60)
```

is translated into:

```
at my-callback@bug1_client, bug1.scm:10
```

The mapping of identifiers is straightforward. It merely uses the standard Hop functions for mangling/demangling identifiers. The Hop mangling is also used to separate stack frames corresponding to Hop function calls and to JavaScript function calls.

Actual source locations are reconstructed by the Hop client runtime system using the extra informations produced by the Hop-to-JavaScript compiler. It relies on the *source map* tables (Lenz and Fitzgerald, 2011) the compiler generates for the JavaScript steppers. These tables contain all the informations needed to map JavaScript source positions into Hop source positions. To make the source tables explicitly available from standard JavaScript code, the Hop-to-JavaScript compiler generates the extra call at the end of each generated file:

```
hop_source_mapping_url( "bug1.js", "bug1.js.map" );
```

This merely registers that a source map table is available for the file `bug1.js`. When a stack frame referring `bug1.js` has to be translated, the table is actually downloaded from the server, and a JavaScript client library translates the JavaScript location translated into a Hop location.

3.3 Enforcing Types and Arity

JavaScript tolerates many user errors. A function can be called with less or extra parameters than required. An unbound variable can be set or an undeclared data structure field accessed without raising errors. The arithmetic operators never raise exceptions whatever the type of the arguments they receive. Many other examples can be found on the Web. As a consequence of this tolerant semantics, Hop cannot delegate to JavaScript the dynamic checking of the programs it compiles. Rather, it must checks types, function arities, and array bounds by itself.

Former studies of the compilation of strict functional languages have shown that for Hop-like languages, despite functions being first class citizen, the compiler knows the very function that is invoked on most call sites (Shivers, 1988; Rozas, 1992). Hence, most arity checks can be resolved statically by the compiler. It is then sensible not to instrument the function bodies which would penalize all the executions of the functions but to instrument only the calls to unknown functions instead. For a call site such as:

```
(fun a0 a1)
```

when the compiler knows that `fun` is a function waiting two parameters, it generates the following JavaScript code:

```
fun( a0, a1 )
```

Otherwise, it generates:

```
hop_check_arity( fun, 2 )( a0, a1 )
```

The arity of the Hop compiled functions is stored in a field called `hop_arity`, the function `hop_check_arity` merely compares the value of this field to the actual number of arguments. For hand-written JavaScript functions the field does not exist and the test always succeeds. The library function `hop_check_arity` is implemented as follows (for simplicity only fixed arity functions are considered in the paper):

```
function hop_check_arity( fun, arity ) {
   if( "hop_arity" in fun ) {
      if( fun.hop_arity == arity ) {
         return fun;
      } else {
         throw new ArityError( fun, arity );
      }
   } else { /* a plain JavaScript function */
      return fun;
   }
}
```

Checking the types and the array bounds is a different matter as in general actual types and array bounds are unknown at compile-time. Hence, before accessing any data structure or any array, a check must be executed. To avoid code size expansion of the generated JavaScript codes, Hop instruments the getters and setters defined in its standard library instead of adding extra tests in the user code it generates.

Checking types at runtime slows down the execution for two reasons. First, it requires to fetch the dynamic types of the objects and compare them against

global values. Second, it prevents the compiler from inlining operators which contain type tests as it would enlarge generated code size too much.

Fortunately, these costs can be significantly reduced using JavaScript dynamic properties which are available since ECMAScript 5. A dynamic property is syntactically used as a regular property but the actual access goes through a user defined function. Let us illustrate how Hop uses this feature with pairs objects which are used to represent list elements. In production mode, pairs are defined as:

```
function sc_Pair(car, cdr) {
   this.__hop_car = car;
   this.__hop_cdr = cdr;
}
```

In debugging mode, the implementation is changed for:

```
function sc_Pair(car, cdr) {
   this._safe__hop_car = car;
   this._safe__hop_cdr = cdr;
}
Object.defineProperty( Object.prototype, "__hop_car", {
   /* type errors for non pair objects */
   get: function() {typeError("car", "pair", this);},
   set: function(v) {typeError("set-car!", "pair", this);}
} );
Object.defineProperty( sc_Pair.prototype, "__hop_car", {
   get: function() { return this._safe__hop_car; },
   set: function(v) { this._safe__hop_car = v; }
} );
```

Using JavaScript getters lets the compiler generate the same inlined code for accessing the first and second elements of pairs. In debugging mode, these accesses now go through the JavaScript *getters* and *setters* which transparently enforce type consistency. This technique has several advantages over inserting extra type checks in the generated code. First, the code of the compiler is unchanged. Second, the generated code is no larger in debug mode than in production mode. Third, even pairs accessed from JavaScript are type checked.

We have evaluated the performance of this technique on four browsers: Firefox 23, Chromium 29, Safari 7, and IE11. Firefox and Chromium have been executed on the GNU-Linux 3.10.10 running on an Intel Xeon E5-1660, 3.3GHz. Safari has been executed on a 2.6GHz core i7 running MacOS 10.9.1, and IE on a core 2.7GHz i5 running Windows 7. We have compared the performance of unsafe access, explicitly type check, and property verified access[3].

[3]Note that it would make few sense to compare the raw performances of the browsers as the hardware running them differs.

	unsafe	check	property
Chromium	44ms	209ms	44ms
Firefox	45ms	177ms	296ms
Safari	39ms	665ms	4678ms
IE	91ms	458ms	1305ms

This experiment shows a contradictory result. On the one hand, there is a huge benefit in using dynamic properties instead of using explicit type checks on Chromium. On the other hand, there is a drawback in using them on other browsers. As the techniques for implementing properties efficiently is now well known because published (Schneider, 2012), we foresee that all browsers will implement them efficiently too in a close future. Since properties have the advantages of their own mentioned earlier we have opted for using them, although they currently penalize non V8-based browsers.

Type checking numerical operators is also required in debugging mode as JavaScript replaces silently all non numeral values with `0` or `NaN`. For that, a prelude is added to each operation which enforces the types of the arguments. For instance, the '`-`' function is defined as:

```
function sc_minux2( x, y ) {
   if( typeof x !== "number" )
      sc_typeError( "-", "number", x, 3 );
   if( typeof y !== "number" )
      sc_typeError( "-", "number", y, 3 );
   return x - y;
}
```

This slows down dramatically the performance because it adds two extra tests and it disables inlining for avoiding code size explosion. We have tried to improved the implementation by testing only the result of the operation as:

```
function sc_minux2( x, y ) {
   var res = x - y;
   if( isNaN( res ) ) {
      if( typeof x !== "number" )
         sc_typeError( "-", "number", x, 3 );
      else
         sc_typeError( "-", "number", y, 3 );
   }
   return res;
}
```

The JavaScript function `isNaN` is a property of the global object. It is then difficult for a compiler, to implement it as efficiently as a simple operator. As suggested by the ECMA-262-5 specification (ECMA, 2009) (see Section 15.1.2.4), calling `isNaN(`*v*`)` can be replaced with the more efficient expression *v*`!==`*v*. We have compared the three methods on the `fib` benchmark which uses extensively additions, subtractions, and integer comparisons.

	type-*based*	isNaN-*based*	!==-*based*
Chromium	71ms	101ms	71ms
Firefox	19ms	101ms	54ms
Safari	48ms	100ms	55ms
IE	129ms	261ms	136ms

This experiment has contradicted our intuition as it clearly shows the benefit of using the explicit type checking approach on all browsers, even if it executes more tests at runtime.

3.4 Global Performance Evaluation

For constructing the contextual stack, an exception is raised each time a callback is registered. Whatever the speed of raising and catching JavaScript exceptions, this has a minor impact on the overall performance as callbacks are installed when pure JavaScript performance does not matter: the network times dominate the preparation of the evaluation of a `with-hop` call; the setup time of a timeout is unimportant as its purpose is to slow down the execution or to yield the processor. The parsing of the string denoting the execution stack and the mapping the JavaScript source locations to the actual Hop source locations would have a much significant impact as it allocates many temporary data structures. This takes time and this exercises the garbage collector. Hopefully it can be delayed until a stack must be actually displayed. In the fast path where no error is raised, the parsing of the stack is never executed.

Performance of arity checking and type checking is a different matter as these require extra tests that get executed frequently. We have measured their impact on a set of representative benchmarks. As we have already established that the performances of Hop client-side programs is on par with similar hand-written JavaScript programs (Loitsch and Serrano, 2007), the results presented here also give an intuition of how JavaScript would be impacted if it was supporting an equivalent debug mode.

For each benchmark, we have measured the raw speed obtain in production mode, the speed of the instrumented version where only the arity is checked, and the fully instrumented version where types are also systematically checked. The results of this experiment are presented Figure 1.

Measuring the performance of browser-hosted JavaScript execution engines is not easy. Profiling tools are unusable as they introduce a bias of their own. Measuring the CPU or system time is impossible as the JavaScript library lacks accesses to the system information. The only possibility we have consists in measuring the user observable execution times. To make performance evaluation even more complex, JavaScript usually stops programs that run more than 5 seconds. We then have to measure short lasting executions, which is subject to a lot of variations depending on the system state (cache, JIT information, etc.). In this context, we have found relevant to run a benchmark several times, collecting all the execution times. When the sum of these times reaches or exceeds 10 seconds, the benchmark is stopped and the mean time is reported. This methodology is not totally accurate but it gives a decently correct idea of the actual raw speed of a benchmark.

The first observation is that degradation varies significantly from one benchmark to another but on Chromium and Firefox, the two most popular portable browsers, we deem it acceptable. In general, Chromium is the less impacted browser. This probably is a consequence of its dynamic properties implementation observed in Section 3.3 that lets it enforce types efficiently.

The second observation is that degradations seem not to depend on the benchmark themselves. For instance, in debug mode, Fib behaves poorly on Chromium and decently on Firefox. An opposite situation is observed for Earley. Probably, for a reason or another, enforcing types on these benchmarks defeats the JIT compilers differently. As we consider a worse ratio of 4 or 5 for debugging acceptable, we have not investigated any further in order to minimize the impact of type checking on performance.

The third observation is that arity checking is unnoticeable but on the two benchmarks Beval and Conform. These two benchmarks use higher order functions extensively. This programming style imposes arity verification at runtime. The overhead is then expected.

4 RELATED WORK

Multitier programming for the Web has been pioneered by GWT from Google, Links from the University of Edinburgh (Cooper et al., 2006), and Hop. The three languages have appeared almost simultaneously in 2006. Other languages have then followed such as Ocsigen (Vouillon and Balat, 2013), iTask3, or Opa (Binsztok et al., 2013). Among the Hop competitors, only GWT considers the problem of debugging Web applications. GWT supports debugging of multitier applications but cannot debug JavaScript components. GWT has nothing similar to the Hop multitier stack.

Nodejs is a platform built on top of V8, the JavaScript runtime used by Chrome. Nodejs is used for building fast, scalable network applications, such

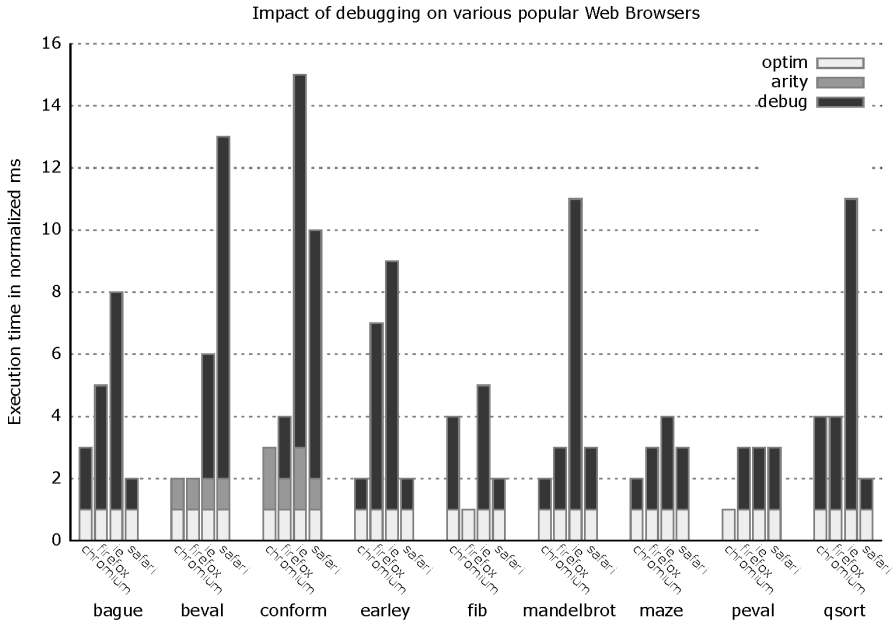


Figure 1: This experiment reports about the impact of debugging on client-side speed. For each benchmark and for each browser, the optimized speed is used as a base value for that benchmark. The experiment presents the slowdowns imposed by the debugging mode. For instance, on the Bague benchmark, arity checking incurs no performance penalty to Chromium and Firefox but type checking slows down Chromium by a factor of 3 and Firefox by a factor of 5.

as Web servers. Nodejs is an effective way of supporting JavaScript on both ends of the application. However a Nodejs Web application is still conceived as two separated software components and debugging is also separated.

The following example mimics the server error example of Section 2.2.

```
var http = require("http");
var url = require("url");

http.createServer(function(request, response) {
  var url_parts = url.parse(request.url, true);
  var query = url_parts.query;

  response.writeHead(200, {"Content-Type": "text/html"});
  response.write("<html>"+query["x"]["car"]+"</html>");
  response.end();
}).listen(8888);
```

When executed, it produces the following trace.

```
node1.js:10
               + query["x"]["car"]
                            ^

TypeError: Cannot read property 'car' of undefined
 at Server.<anonymous> (node1.js:10:40)
 at Server.EventEmitter.emit (events.js:98:17)
 at HTTPParser.parser.onIncoming (http.js:2056:12)
 at HTTPParser.parserOnHeadersComplete (http.js:120:23)
 at Socket.socket.ondata (http.js:1946:22)
 at TCP.onread (net.js:525:27)
```

The error is correctly located in the server source file but the stack trace is oblivious of the client-side execution that has preceded the server computation. It merely reports that the error as occurred in the context of answering an HTTP request but without much details. Running Nodejs in debug mode could give access to extra informations about the nature of the HTTP request but it will still lack informations about the client state. The techniques proposed in this paper could improve this situation.

Popular modern JavaScript frameworks raise the abstraction level of client-side programs by offering facilities for generating client-side programs at runtime and for communicating with the server more easily. This makes developing applications easier but as of the current versions, it also makes debugging more difficult because the code automatically generated by the runtime system shows up when an error is raised. Let us illustrates this with the Google's Angularjs framework (Google, 2013). Let us consider the tutorial available on the framework Web page which illustrates Ajax programming with the following example:

```
var catApp = angular.module('catApp', []);
catApp.controller('PhoneListCtrl', ['$scope', '$http',
   function ($scope, $http) {
      $http.get('phones.json').success( function(data) {
      $scope.phones = data; });
  $scope.orderProp = 'age'; }]);
```

Introducing a syntax error in `phones.json` produces:

```
SyntaxError: Unexpected token }
    at Object.parse (native)
    at fromJson (angular.js:1035:14)
    at $HttpProvider.defaults.defaults.transformResponse
       (angular.js:6926:18)
    at angular.js:6901:12
    at Array.forEach (native)
    at forEach (angular.js:302:11)
    at transformData (angular.js:6900:3)
    at transformResponse (angular.js:7570:17)
    at wrappedCallback (angular.js:10905:81)
    at angular.js:10991:26
```

As it can be seen, the whole execution trace is only populated with Angularjs entries which none is explicitly mentioned in the user program. Even more important, the stack trace it totally silent about the `controller`'s code and the HTTP request. The techniques presented in this paper could help presenting less obscure stack traces to the programmer.

The tolerant semantics of JavaScript makes it inconvenient to be used as the target of the compiler of another programming language. The checks executed at runtime by JavaScript are likely to be either insufficient if the source language is safe, or superfluous if the source language is unsafe. In Hop debugging mode, each access is double-checked. Once by Hop and once by JavaScript which ensures that the execution does not corrupt the memory. Of course, this double checking slows down executions. In this respect, the `asm.js` endeavor (Herman et al., 2013) is promising. It could become an interesting target for the Hop client-side compiler.

There is a whole line of research which consists in typing JavaScript. Some focus on inferring static types of JavaScript programs (Jensen et al., 2009), some such as TypeScript (Microsoft, 2013) extend the language to support type annotations. The shared objective is to enable JavaScript errors detection at compile-time. This is orthogonal to our effort as our purpose is to detect unexpected behaviors at runtime.

5 CONCLUSIONS & PERSPECTIVES

The lack of complete debugging information is acknowledged as a major difficulty when developing Web applications (Schrock, 2009). Using the Hop multitier setting we have solved this problem by creating a debugger which reports full stack traces. When an error is raised, the programmer is presented with the complete execution trace composed of server-side and client-side frames that have conducted to the error. The Hop runtime environment supports a debugging mode where types, arity, and bounds are strictly enforced. Combining the debugger and the debug mode makes error easier to localize and to understand.

The presented debugger exposes a unified execution stack that reflects both ends of the application but it uses two separate steppers that cannot collaborate. In a further step, we will create a global stepper that will be able to traverse the network. Stepping forward seems easy to obtain because it will just require the implementation of a collaboration layer between two existing tools. Inspecting the execution stack backward is more hypothetical since it requires to save execution traces potentially infinitely.

The presented techniques rely on the multitier paradigm to expose a global and coherent view of the execution between the server and the client. They also rely on code generation to instrument the code actually executed on the browser. In Hop this is implemented in a single runtime environment whose main element is a custom bootstrapped web server embedding compilers for generating HTML and JavaScript on-the-fly. In addition to supporting better debugging, this approach also enables fast dynamic HTTP responses servers (Serrano, 2009). This approach also has drawbacks: as it is does not rely on mainstream tools and techniques, it attracts few developers. In consequence, Hop offers less libraries and support than popular languages. Our response to this problem consists in supporting a deep compatibility between Hop and the standard Web technologies in order to make HTML, CSS, third party JavaScript programs, and Web services as easy to use in Hop as they are in JavaScript, PHP, or any other standard tool (Serrano and Berry, 2012). It remains that Hop relying on the Scheme programming language, an academic programming language, is unlikely to get massive adoption. As this is our objective, we now consider a radical evolution. We are then considering transposing the Hop principles to JavaScript. The recent evolutions of JavaScript such as the EcmaScript 6 quasi-literal extensions (ECMA, 2013) that offers a standard mean for expressing client- and server-side code within a single JavaScript source file and the fast V8-based server-side environments are paving the road to new Web development environments for which the contributions presented in this paper might be useful.

REFERENCES

Binsztok, H., Koprowski, A., and Swarczewskaja, I. (2013). *Opa: Up and Running*. O'Reilly Media.

Boudol, G., Luo, Z., Rezk, T., and Serrano, M. (2012). Rea-

soning about Web Applications: An Operational Semantics for HOP. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 34(2).

Cooper, E., Lindley, S., Wadler, P., and Yallop, J. (2006). Links: Web Programming Without Tiers. In *5th International Symposium on Formal Methods for Components and Objects (FMCO)*, pages 266–296, Amsterdam, The Netherlands.

ECMA (2009). Ecma-262: Ecmascript language specification.

ECMA (2013). Ecma-262: Ecmascript quasi-literals (ecma-262 edition 6, rev 21). http://wiki.ecmascript-.org/doku.php?id=harmony:quasis.

Feldman, P. (2011). WebKit Remote Debugging. http://www.webkit.org/blog/1620/webkit-remote-debugging.

Google (2013). Angularjs. http://angularjs.org/.

Herman, D., Wagner, L., and Zakai, A. (2013). asm.js, Working Draft - 17 march 2013. http://asmjs.org/spec-/latest/.

Jensen, S., Møller, A., and Thiemann, P. (2009). Type analysis for javascript. In *Proceedings of the 16th International Symposium on Static Analysis (SAS)*, pages 238–255, Berlin, Heidelberg. Springer-Verlag.

Kellomäki, P. (1993). Psda portable scheme debugger. *SIGPLAN Lisp Pointers*, VI(1):15–23.

Lenz, J. and Fitzgerald, N. (2011). Source Map Revision 3 Proposal – better bidirectional mapping.

Loitsch, F. and Serrano, M. (2007). Hop Client-Side Compilation. In *Proceedings of the 8th Symposium on Trends on Functional Languages (TFL)*, New York, USA.

Mickens, J. (2012). Rivet: Browser-agnostic Remote Debugging for Web Applications. In *Proceedings of the Usenix Annual Technical Conference (ATC)*, pages 30–43, Boston, MA, USA.

Microsoft (2013). TypeSscript, Language Specification, version 0.9.5.

Rozas, G. (1992). Taming the Y operator. In *Conference Record of the 1982 ACM Symposium on Lisp and Functional Programming (LFP)*, pages 226–234, San Francisco, California, USA.

Schneider, F. (2012). High performance JavaScript with V8. http://cs.au.dk/ jmi/VM/IC-V8.pdf.

Schrock, E. (2009). Debugging AJAX in Production. *ACM Queue*, 7(1).

Serrano, M. (2009). HOP, a Fast Server for the Diffuse Web. In *proceedings of the 11th international conference on Coordination Models and Languages (COORDINATION'09)* (invited paper), Lisbon, Portugal.

Serrano, M. and Berry, G. (2012). Multitier Programming in Hop - a first step toward programming 21st-century applications. *Communications of the ACM*, 55(8):53–59.

Serrano, M., Gallesio, E., and Loitsch, F. (2006). HOP, a language for programming the Web 2.0. In *Proceedings of the First Dynamic Languages Symposium (DLS)*, Portland, Oregon, USA.

Sharma, A. J. (2012). Better Web Development with WebKit Remote Debugging. In *Proceedings of the World Wide Web Conference (WWW)*, Lyon, France.

Shivers, O. (1988). Control Flow Analysis in Scheme. In *Proceedings of the SIGPLAN '88* ACM SIGPLAN *Int'l Conference on Programming Language Design and Implementation (PLDI)*, pages 164–174, Atlanta, Georgia, USA.

Vckovski, A. (1998). wshdbg - a Debugger for CGI Applications. In *Proceedings of the 6th conference on Annual Tcl/Tk Workshop (TCLTK)*, pages 89–96, Berkeley, CA, USA.

Vouillon, J. and Balat, V. (2013). From bytecode to Javascript: the Js_of_ocaml compiler. *Software: Practice and Experience*, doi: 10.1002/spe.2187.

Authentication and Authorisation for Widget-based Applications in a Loosely-coupled Web eLearning Environment*

Jean-Noël Colin and Minh Tien Hoang
PReCISE Research Center, University of Namur, Namur, Belgium
{*jean-noel.colin, minh-tien.hoang*}*@unamur.be*

Keywords: Widgets, eLearning, Authentication, oAuth, OpenId.

Abstract: In this paper, we discuss the mechanisms used for authentication and authorisation of users in a widget-based web environment that integrates multiple components from multiple providers running at multiple locations. Users are typically teachers and pupils who access our platform either directly or through their usual Learning Management System (LMS). Although the technologies used themselves are not new, we believe that the approach we follow is innovative enough in their integration and use. Our approach is based on a survey ran among european teachers about their practices in terms of user credentials usage and sharing.

1 INTRODUCTION

The last few years have seen the emergence of a new approach to building Web applications, in the form of mashups. Mashups are applications that reuse and combine data and services available on the web (Aghaee and Pautasso, 2012). In the rest of the paper, we will follow the definition of mashup and related concepts as defined in (Zibuschka et al., 2010). Web mashups usually combine data and services that are available on the Internet– freely, commercially, or through other partnership agreements. In this perspective, we differentiate between the mashup platform (the system that integrates remote content) and the remote content itself, usually integrated through some standard protocol or approach like W3C Widgets [2].

In terms of security, one of the main issues encountered when dealing with complex software web architecture is that the traditional browser security model dictates that content from different origins cannot interact with each other, while content from the same origin can interact without constraint (De Keukelaere et al., 2008). While solutions exists (like using a proxy server or making use of `script` tags), we have to find a generic way to overcome this limitation and allow content from mixed origins to interact with each other.

A key element to this integration is the proper management of users, i.e. know who is who, and who is allowed to do what: authentication and authorisation. Users may have an account with the mashup platform or with the provider of the remote service or content; moreover, users may want to use credentials they already have with other actors on the Internet (think of the 'Sign-In with your *xxx* account' button you find on many websites). Access policies may be defined centrally, while others may be specified on the remote side. This is what we call a loosely-coupled environment: independent components interacting together through identified interfaces, with little to no knowledge about remote components and policies. The challenges we face in this work are the following:

- how do we solve user authentication in such multi-layer loosely-coupled environment?
- how do we ensure proper access to mashup services and remote content, based on a decentralised policy
- how do we propagate authentication and authorisation information through the various layers of the environment?

This research takes place in the framework of the

*The work presented in this paper is partially supported by the European Commission's FP7 programme –project iTEC: Innovative Technologies for an Engaging Classroom (Grant agreement N°257566). The content of this paper is the sole responsibility of the authors and it does not represent the opinion of the European Commission and the Commission is not responsible for any use that might be made of information contained herein.

[2]http://www.w3.org/TR/2012/REC-widgets-20121127/#widgets-family-of-specifications

the iTEC project[3], a 4 years EU-funded project focused on the design and evaluation of the future classroom in Europe.

To solve the various issues and to ensure interoperability, we have investigated several open standards and solution, and have selected oAuth and OpenID as building blocks of our solution, in an approach similar to the one presented in (Govaerts et al., 2011). We will discuss this choice further in the paper, as well as give a complete description of the solution we put in place. The rest of the paper is organised as follows: section 2 presents the conclusions of a survey run with teachers to identify use of Learning Management Systems and practices regarding use and re-use of credentials. The results of the survey guided the rest of the work. Section 3 describes the overall project architecture and requirements in terms of user management and access control. Section 4 then presents the solution design and implementation. Section 5 gives some concrete examples. Finally, section 6 provides some conclusions and ideas for future work.

The paper does not include a related work section, because although it relies on solid and proven components and approaches, we could find no publication relating to a similar approach of building a centrally-managed environment for authentication and authorising access to services in a loosely-coupled web environment while accepting external authenticators. This may be linked to the highly decoupled and distributed approach which is inherent of large European research projects.

2 LESSONS FROM THE TEACHERS SURVEY

A survey was run during the second year of the project (2012) to ask teachers about their current practices regarding the use of Learning Management Systems, the type of credentials they use, their familiarity with third parties that provide authentication services (Google, Facebook, Microsoft Hotmail, Yahoo) as well as their inclination towards re-using credentials among various websites.

We received a total of 269 responses from 17 countries. Distribution across countries is unbalanced, but since the survey is only indicative, this is not considered as a problem.

Results showed that Moodle is the most widely used LMS, but quite a long list of other systems are also in use. The LMS is installed either locally to the school, or shared among different schools, although almost half of the respondents did not really know.

More importantly, responses showed that although a majority of users have local credentials to login to their LMS, a significant number of them use external credentials, which can be provided by an education authority (school, regional or national authority) or by third parties (such as Google, Facebook, etc.)

Finally, the survey showed that the vast majority of teachers (and students) have accounts with one of the major Identity Providers on the Web (Facebook, Google, Yahoo) and that almost 70% of them are willing to re-use those credentials to access their school services.

However, it is important to note that integration with external Identity Providers is not always an option, due to technical restrictions defined at the school level (site-blocking firewall).

Our conclusion is that if we want to add new services to existing LMS already made available to schools, we shall have to integrate them without adding an extra authentication burden on users. It means that new services will have to extend existing infrastructure and offer the possibility of re-using credentials that users may already possess. However, because some users are concerned that re-using credentials might constitute a security risk, it is important to propose a mixed approach (i.e. use both siloed and re-used credentials).

Complete results of the survey are presented in (Colin and Simon, 2012)

3 PROJECT DESCRIPTION AND ARCHITECTURE

The iTEC project (*i*nnovative *T*echnologies for an *E*ngaging *C*lassroom) is a four-year project funded by the EU; it is focused on the design of the future classroom in Europe. The project, which involves 27 partners among which 15 Ministries of Education from across Europe, brings together teachers, policymakers, pedagogical experts - representatives from each stage of the educational processes - to introduce innovative teaching practices.

The goal of iTEC is to pilot learning and teaching scenarios using integrated technology solutions in over 1,000 classrooms in 15 countries, making it by some margin the largest pan-European validation of ICT in schools yet undertaken. The scope of the project ranges from pedagogical scenario design, customisation to local technical environments, development of required technologies and school piloting. In this paper, we will focus on the technical aspects only.

[3]http://itec.eun.org/

3.1 Basic Concepts

The idea behind iTEC technical implementation is to pull pedagogical content and tools together into a common interface in a meaningful way, i.e. so that they can support a specific pedagogical scenario. The basic building blocks of the iTEC ecosystem (aka the iTEC Cloud) are the following:

- a *shell* is the host platform (in mashup terminology, this is the Mashup-providing platform) used to integrate the various forms of content and tools; it acts as an empty 'nest' for those remote components. It is also the first user-facing component. No particular assumption is made on the shell technology, beyond the fact that it must be able to display widgets (Wilson et al., 2011). Examples of shell platforms are Learning Management Systems (LMS) like Moodle[4], mobile platform like Android[5] or Interactive Whiteboard software like OpenSankore[6]
- a *Web application* is an autonomous piece of code that offers a set of a self-contained services. It takes care of its own rendering and embeds all necessary services to serve end-users and fulfill expected functionalities. It may include user management functionalities or rely on third-party for that purpose.
- A *widget* is an elementary piece of application or content (image, video...); it embeds some functionality that is of interest to the users and is displayed through a shell. A widget does not operate on its own, but rather, requires a widget engine or runtime to create and manage widget instances and a shell to be made available to users. Several kinds of widget engines exist, like Yahoo Konfabulator, Apple's Dashboard, Opera Widgets runtime, Apache's Wookie server. A self-contained widget is a widget that embeds all its functionalities, like a calculator widget. A WebService-backed widget is a widget that relies on other components to function. Such component may take the form of backend services reached through a REST webservices interface. This would be the case for a chat widget that stores chat messages on a separate distant datastore. Widgets may offer collaboration services, like chat or videoconferencing, access to pool of resources, geolocation functionalities or administrative functions (user management...)
- A *backend service* is a software component with no user interface, that is accessed by other components through a web service interface. In iTEC, most components are built in such a way that the user interface is implemented by a widget while the business logic and data layer are implemented by a backend service, although in some cases, components are developed as a standalone web application, with or without access to a backend service.

iTEC's approach to assemble and deliver content and tools is through the integration of widgets in a shell; the combination of widgets is derived from a pedagogical scenario, with the aim of supporting it by providing the right tools, which may be synchronous or asynchronous collaboration and communication, calculator, simulator, geolocation or content sharing. In addition to widgets, iTEC also uses web applications for some of its services.

3.2 iTEC in Action

Starting from a pedagogical scenario, widgets are assembled in a shell's space, typically a course page in a LMS or a presentation slide in an Interactive Whiteboard notebook, in a way that they can be used by end-users, a teacher and his class for example. To support this type of usecase, iTEC designed a set of components:

- the *widget store* is the main repository for widgets. It stores widget descriptions, allowing for searching, tagging, rating and commenting about widgets. It is similar in its concept to Google's PlayStore or Apple's AppStore. It is built as a widget.
- the *Composer* is used to manage pedagogical scenarios, and assemble widgets that meet the requirements of those scenarios. For instance, a scenario may include synchronous communication, picture sharing and collaborative writing, and this would translate into a chat-, a camera- and a wiki widget. This instantiation takes into account the technical settings of the target environment, i.e. the set of technologies available in a specific school. The Composer is built as a web-application.
- the *People and Event directory* is a database of people and events that may be of interest to pedagogical actors; it can be a subject expert, or a conference or similar event. Those resources are also rated, tagged and commented. This component is built as a web-application with an integrated backend service, thus supporting the development of

[4] https://moodle.org
[5] http://www.android.com
[6] http://open-sankore.org

widgets.

- the *Scenario Development Environment* is a recommendation engine, that can collects data from the above components, and provides recommendation to the Composer in its scenario instantiating role, by suggesting most appropriate resources to include in the setup. This component is developed as a pure back-end service since it is not meant to support any type of user interaction.

3.3 Authentication and Authorisation Challenges

From the description above, iTEC integrates a wide variety of components: shells, web applications, self-contained widgets, widget-based applications... This integration raises some questions in terms of user management and access control:

- user authentication may take place at the shell level, but also, some integrated services may require some form of authentication or at least be aware of the visiting user's identity. This implies the need for an authentication mechanism that can span the range of components and provide a consistent information about the user
- access control policies may be defined centrally, at the iTEC Cloud level, but these policies have to co-exist and be consistent with those defined at the shell level, or at the integrated services level, if any. Again, this requires an authorisation mechanism that integrates at the various levels of the architecture.

Our goal is thus to design a system that would meet the following requirements:

- allow user authentication at the shell level, and pass the information into sub-components (widgets and back-end services)
- allow access policies to be defined globally to the iTEC Cloud, based on a Role-Based Access Control (Ferraiolo et al., 2001) model, and have those policies propagated to the sub-components
- from the global access rules, provision local policies at the level of iTEC back-end services
- designed solution should support interoperability with major service providers

4 PROPOSED SOLUTION

The interoperability requirements led us to focus on open standards and protocols to build authentication and authorisation mechanisms. We performed a thorough study, and identified candidate protocols like SAMLv2[7], OpenID[8] and oAuth[9]. Due to their technological maturity, their relative simplicity, their support for web interactions, the availability of libraries and their wide adoption by main actors on the net, we selected oAuthv2 and OpenIDv2 as the basis for our solution. The fact that users are warned when an application wants to access protected data was also an element of choice.

OpenIDv2 (Foundation, 2007) is an open and standard protocol for signing on to websites using one single set of credentials. The protocol has been developed for many years and adopted by major players on the Internet, like Google. It relies on the assumption that users have an identity defined with an Identity Provider (IdP), and want to use that identity to access various services offered by Service Providers (SP). The typical flow is a user visiting a Service Provider that requires authentication; SP prompts the user for her identity or that of her IdP. The user is then redirected to the IdP to authenticate, and if authentication succeeds, the user is sent back to the SP with the proof that successful authentication did take place. Optionally, the IdP may provide additional information about the user (this requires some protocol extensions).

oAuthv2 (Hardt, 2012) is a protocol for managing delegation of authorisation. Its main use case is a user (the *resource owner*) needing to give access to some of its resources hosted on a server (the *resource server*) to a *client*, typically another service. To avoid forcing the user to give her credentials to the client, oAuthv2 introduces a workflow where when the user is asked by the client to give access to a resource, she is sent back to an *authorisation server* where she authenticates and is then asked to grant or deny access. Upon success, the authorisation server issues an *access token* to the client that it will use to access the resource on behalf of the user. In this way, the user's credentials are never disclosed to the client. This is the protocol that Facebook or Yahoo use for granting access to their services to remote sites, after getting the agreement of the user. oAuthv2 supports various types of 'grants', to support different profiles of this protocol and accommodate different situations:

Authorisation Code Grant this is the most secure scenario, in which the client directs the resource owner to the authorisation server for authentication and access request; upon success, the authorisation server issues an *authorisation code* to

[7] http://saml.xml.org/
[8] http://openid.net/
[9] http://oauth.net/2/

the client, that the client then exchanges with the authorisation server for an *access token*, that is finally presented by the client to the resource server to get access to the resource. All interactions with the resource owner go through her user-agent (typically her browser). This scenario supports client authentication by the authorisation server before issuing an access token, and ensures that the access token never reaches the resource owner's user-agent, which could lead to token leakage.

Implicit Grant this is a simplified version of the previous scenario, in which instead of being issued an authentication code by the authorisation server, the client directly receives an access token. This scenario is targeted at clients implemented in a browser, typically in javascript. In this case, the authorisation server does not authenticate the client, and the access token is exposed to the resource owner or other applications with access to its user-agent.

Resource Owner Password Credentials Grant this scenario is built on the assumption that there exists a high degree of trust between the resource owner and the client. The resource owner provides the client with her credentials, and the client uses them to request an access token from the authorisation server. This scenario supports client authentication.

Client Credentials Grant in this scenario, the client is acting on its own behalf, not on behalf of the user. The client authenticates directly to the authorisation server and receives an access token.

It is worthwhile noting that oAuthv2 also supports extension grants that allow to extend the token request mechanism to support different types of credentials, like SAML assertions.

Because we chose to use oAuthv2 to secure widget access to back-end services, and because widgets usually involve client-side computing and get access to the user's environment, the implicit grant is the only option of choice. However, we also successfully implemented the client credentials grant to secure access to the SDE backend service. One of the drawbacks of the implicit grant is the absence of client authentication, but this can be explained by the nature of widgets, which are running client-side, making available any sensitive information to other components running in the user's environment (user-agent). It would thus not be possible to securely store client credentials at the widget level.

To integrate those protocols into the iTEC environment, we designed the UMAC (User Management and Access Control) framework, which comprises the following components:

- the *UMAC server* is responsible for user authentication, issuance of tokens, and management of user data and privileges; it plays the role of the OpenID's Identity Provider, the oAuth's authorisation server, and implements a back-end service to access, store and manage user data and privilege information.
- the *UMAC filter* is an authorisation guard that sits in front of back-end services; the back-end service represents the oAuth's Resource Server, and the UMAC filter is in charge of validating access tokens.
- the *UMAC widgets* are a collection of widgets that allow to access and manage authentication and authorisation information in the iTEC Cloud. Those widgets allow to register a new user, to update a user's details, to create sets of users, and to assign iTEC roles.
- the *UMAC library* is a JavaScript library of tools to help the widget developer to easily integrate with the UMAC framework and not care about the various protocols' implementation.

4.1 UMAC Server

The UMAC Server serves two main purposes: authenticating users and controlling access to back-end services.

To authenticate users, UMAC Server implements the OpenID Provider specification. It handles authentication requests from iTEC relying parties, typically shells or web applications, authenticates users, and responds to relying parties; UMAC Server supports SREGv1.0 and AXv1.0 OpenID extensions to provide basic information of logged in user (username, first and last names, email address, language, timezone, country). Authentication is checked against a local database of users.

One of the requirements drawn from the survey described in section 2 mandates that iTEC should allow users to login using third-party credentials, namely Google, Facebook or Yahoo. Thus the UMAC Server supports user authentication using any of those systems, by implementing an OpenID Relying Party (in the case of Google and Yahoo) and an oAuth client (in the case of Facebook).

Access control to iTEC services is handled by UMAC Server. Access requests may come from widgets or web applications, in which case the oAuthv2 scenario implemented is the implicit grant, but requests may also come from standalone applications,

which are run in a more controlled environment, and for which the selected scenario is the client credentials grant. Thus UMAC Server implements the related sections of the oAuthv2 specification, and handles Authorisation Requests (for the implicit grant) and Access Token Requests (for the client credentials grant), issuing access tokens to widgets and controlled applications respectively. A token is a random string concatenated with a timestamp.

In addition to the authentication and authorisation functionalities, UMAC server is also used to store user information; this information is made accessible through a REST API, which is protected by the oAuthv2 protocol, just like any other iTEC back-end service. Basic CRUD functionalities are implemented to add, update, delete or get information about a user account. This API is accessed typically by the UMAC widgets.

Finally, UMAC server is used to manage user privileges; those privileges span all iTEC services, i.e. apply equally to shells, widgets or back-end services. Six levels of privileges are defined in a strictly hierarchical way: super-user, administrator, coordinator, teacher, student and guest. The level of privilege of a user is passed to the OpenID relying party upon authentication through SREG or AX extensions, where available, and they are checked by the token validation process between the UMAC filter and the UMAC server.

For a seamless user experience, UMAC authentication is propagated to the shell through a plugin mechanism which is dependent on the shell itself. In this way, once the user is authenticated, all shell components (typically widgets) can reuse the user information.

4.2 UMAC Filter

The UMAC filter is designed to be put in front of back-end services, and interacts with the UMAC server following the oAuthv2 protocol to control access to the services by ensuring that only authorised requests get served. The current implementation of the filter takes the form of a servlet filter, which makes it very easy to integrate and (de)activate and realises a separation of concerns by allowing the service developer to work independently from the access control mechanism.

In the oAuthv2 terminology, the UMAC filter acts as the protection part of the resource server. It receives requests for access in the form of REST calls (basically http requests), and for each requests, it checks that an access token is provided. If no token is present, an error is returned, and it is up to the client to obtain one. If a token is present, its validity is checked by querying the UMAC server through a secure channel, and upon success, the lifetime of the token and the userid of the token owner are returned to the filter. Based on this information, the filter then checks the local access policy that defines the rules for accessing the service. These rules are expressed using the Apache Shiro[10] system. If the rules are evaluated positively, access is granted and the request is passed to the service. Otherwise, an error is returned. For efficiency reasons, the UMAC filter caches the validated tokens for a period of time to avoid unnecessary roundtrips with the UMAC server.

4.3 UMAC Library

The UMAC library is a Javascript library of functions that aims at facilitating the development of widgets and their integration with UMAC authentication service, more precisely, the oAuth authentication endpoint's service. It hides the complexity of the protocol by providing methods to manage the whole authentication process (request for token, redirect to authentication form, token transfer to requesting component and error handling).

Figure 1 presents the UMAC components (in gray) as well as the interactions with other iTEC systems. UMAC Server is used for authentication (solid lines) either from shell, widgets or web applications like the composer. This follows the OpenID protocol. Authentication may be local (using the User DB) or rely on third-party authenticators (right-most box). Regarding authorisation (dashed lines), UMAC widgets allow to register or update user information through the UMAC REST Web Service, which is protected by the UMAC filter. Similarly, any other iTEC component may access iTEC back-end services which are protected by the UMAC filter (see bottom of the diagram).

5 EXAMPLE SCENARIOS

In this section, we provide concrete examples of the system described above.

5.1 Authenticating a User

This is the very first step to do to consume a protected service or piece of content. A user makes an authentication request directly from a widget, from a shell or from a relying web application by clicking on

[10] http://shiro.apache.org/

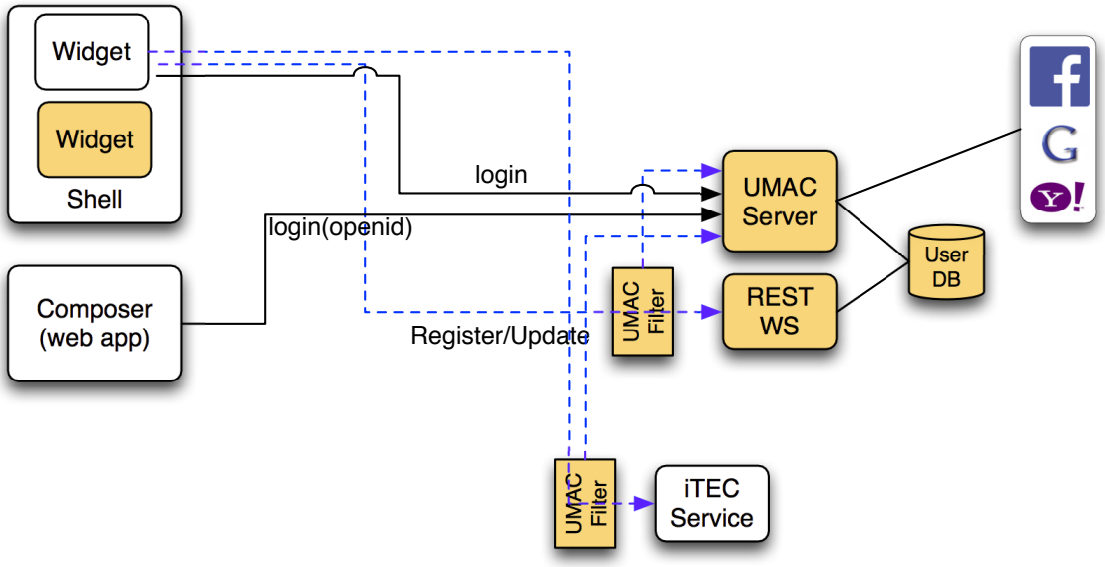


Figure 1: UMAC components.

an URL or a button. This causes redirection to the UMAC Authentication Window (see figure 2),where the user must provide her credentials to UMAC Server or use her external authentication sources such as Yahoo, Google, Facebook to proof his identity. Upon successful authentication, an authentication response (in case of OpenID flow) or an access token (in case of oAuth flow) is generated by UMAC Server and sent back to the Relying Party or the Client respectively.

It is to be noted that even though the authentication action is triggered at the widget level, authentication information is pushed up to the container level (i.e. the Mashup Providing Platform).

5.2 User Authorisation

For fine grain and flexible authorisation, UMAC relies on both a permission-based approach and on a role-based system, in which permissions are assigned to roles rather than being assigned directly to users. Roles are then assigned to users who inherit all permissions linked to the role. At the moment, role structure is flat, i.e. we do not support hierarchical roles.

A permission is basically a tuple that expresses what action can be executed by a service on a specific data scope. Our model follows a Discretionary Access Control approach (Saltzer and Schroeder, 1975) where each piece of data is owned by exactly one user. Permissions are defined according to the specifications below (expressed in Augmented Backus-Naur Form (Crocker, 2008)):

```
permission = action ":" service-name
    *(":" data-scope)
action = (atomic-action *("," atomic-
    action)) / "*"
atomic-action = "create" / "read" / "
    update" / "delete"
service-name = 1*(DIGIT / ALPHA / "-"
    /"\_")
data-scope = 1*(DIGIT / ALPHA) /
    special-data-scope / (data-scope
    *("," data-scope))
special-data-scope = "me" / "mine" /
    "*"
```



Figure 2: UMAC Login window.

The special-data-scope element can take one of

three special values:

- "*me*": this data-scope represents the user currently logged in, i.e. the user who issued the access request; this is typically used for the user management to allow a user to update his own data, but others'
- "*mine*": this data-scope includes all data that is owned by the user currently logged in;
- "*": this data-scope includes all data; it does not need to be specified, e.g.: the permission `action:service-name:*` has exactly the same meaning with `action:service-name`

The next paragraphs provide some concrete examples of expressing UMAC permissions.

The rule "any user can read and update his information and unregister his account" can be decomposed as follows:

- The actions are "*update*", "*read*" (for update and read information) and "*delete*" (for unregistering account).
- The service in used is "*users*".
- The id of the target is the identifier of the requesting user himself, which can be the real id of user as recorded in UMAC, or the special id "*me*".

The rule would then translate to: `update,read,delete:users:me`.

In UMAC this is the default permission assigned to all registered users.

Role-based rules follow a similar approach: for instance, the rule "UMAC admin Role can read and update any user information as well as grant and revoke iTEC Coordinator of Belgium role to a user" includes two different sets of permissions:

- read and update information of any user, which translates to `update,read:users:*` or more briefly `update,read:users`
- grant and revoke belgian coordinator role:, which translates to `read,delete:roles:coordinator:be` (the last `be` data scope represent the belgian record)

Permission to role assignments are stored in the UMAC server database, and are managed through the UMAC widgets; this operation can performed at runtime, which means that the role can be defined dynamically. Moreover, all modification of permission/-role assignments takes effect immediately for all users to whom the role is assigned.

5.3 Securing a Webservice

As stated in section 4.2, backend services are protected by the UMAC filter. Implementation of the filter takes the form of a servlet filter because most iTEC technologies use the Java language and because it is easy to integrate as a proxy in a non-Java environment. As a consequence, the protected web service provider only needs to include the declaration of the UMAC filter in the servlet mapping section of the web.xml configuration file.

In iTEC, the UMAC filter is currently deployed at the People and Event Directory (P&E Directory)[11]. With this deployment, P&E Directory developers only focus on their functional development, all authorisation rules being activated by a simple declaration in its configuration file. Authorisation rules of the filter can be defined dynamically, as shown in section (5.4)

The UMAC filter is also used to protect the UMAC Backend Service (REST WS box on figure 1)

5.4 Central Policy Management

Permissions, roles, role-permissions assignment and user-role assignments are managed by the UMAC Server; they can be updated through the UMAC widgets or any other component that access the UMAC REST WS, provided the appropriate level of privilege is granted.

UMAC supports the `super-admin` role, a static built-in role which is assigned to the UMAC system administrator once it is installed. Since the authentication and authorisation engine is based on Apache Shiro framework(Team, 2013), most of authentication authorisation information are defined in a "shiro.ini" file, including: built-in super admin accounts, default permissions granted to all authenticated user, mapping table between service names and REST service contexts which require authorisation and information for initialising authentication realms (local realm to login locally to UMAC, OpenID realm to login remotely with Yahoo, Google...).

The UMAC filter (section 4.2) has a similar implementation except that there is no authentication database since it relies entirely on the authentication service of UMAC server. It only maintains an authorisation database which can also be configured through REST services. The UMAC filter administrator is defined in "shiro.ini" at filter side; it has similar function to `super-user` at UMAC server but is completely independent. This design helps the filter benefit from authentication service of UMAC server while keeping a totally independent authorisation lattice under the control of the filter administrator.

The UMAC authorisation widgets are developed to manipulate all information related to authorisation

[11] http://itec.eun.org/web/guest/people-and-events

in the iTEC Cloud. The widgets also allow an authorised user to register a backend service provider and assign authorisation rules to that service provider, thus providing a central place to manage all authorisation at the iTEC components level. This allows for a more consistent and secure approach.

5.5 Implementing Widgets and Services

Integrating a new service into UMAC is a very simple process that consists of instantiating the UMAC filter in front of the new service. As stated above, the filter takes the form of a Java servlet, that intercepts all service requests, authenticates them and checks authorisation together with the central policy. A mapping between the centrally-managed iTEC roles and the service-specific roles is defined and translated into the Shire rules.

Widgets access services through the XMLHttpRequest API that allows to connect to remote sites and services. To make the integration with UMAC as easy as possible for widget developer, calls to services are wrapped through calls of the UMAC library that hide the logic of the underlying protocols. From the widget perspective, invoking a service is simply a matter of issuing a request to the service. Authentication and token issuance is handled under the control of the library.

6 CONCLUSION & FUTURE WORK

In this paper, we described a model and its implementation that supports centralised authentication and authorisation in a loosely-coupled multi-layered web application in the eLearning area. The model is open to third-party authenticators, following the conclusion of a survey run among potential users of the platform. Although the individual components of the approach themselves are not particularly innovative, yet state-of-the-art, we advocate that the integration approach we propose, that takes into account requirements and behaviours of end-users, and the complexity of the environment to which it applies constitute the core of the research. It could easily be transposed to other similar environments.

The whole iTEC ecosystem has been used through various cycles of piloting, involving over 2,000 classrooms across 19 European countries, and the UMAC component, although almost invisible to the vast majority of users, did work properly, securing the services from 4 main providers access through a wide variety of widgets.

In the future, we intend to work on a stronger and easier integration between shells and UMAC components, which would allow further customisation of the shell based on information received from the UMAC system. We will explore the IMS Learning Tools Interoperability specifications(Wilson et al., 2011)[12] as it seems to be a potential candidate for supporting tools integration.

A scope should be added to tokens, to reduce the granularity of authorisation rules and limit the potential impact of token interception, and client authentication should be added to the protocol to ensure that only authenticated clients can obtain access tokens from UMAC server.

REFERENCES

Aghaee, S. and Pautasso, C. (2012). An evaluation of mashup tools based on support for heterogeneous mashup components. In Harth, A. and Koch, N., editors, *Current Trends in Web Engineering*, volume 7059 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg.

Colin, J.-N. and Simon, B. (2012). D7.2: Second generation of iTEC shells and composer. Project deliverable 7.2, University of Namur.

Crocker, D. (2008). Augmented BNF for Syntax Specifications: ABNF. RFC 5234, RFC Editor.

De Keukelaere, F., Bhola, S., Steiner, M., Chari, S., and Yoshihama, S. (2008). Smash: secure component model for cross-domain mashups on unmodified browsers. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 535–544, New York, NY, USA. ACM.

Ferraiolo, D. F., Sandhu, R., Gavrila, S., Kuhn, D. R., and Chandramouli, R. (2001). Proposed nist standard for role-based access control. *ACM Trans. Inf. Syst. Secur.*, 4(3):224–274.

Foundation, O. (2007). Openid authentication 2.0. `http://openid.net/developers/specs/`.

Govaerts, S., Verbert, K., Dahrendorf, D., Ullrich, C., Schmidt, M., Werkle, M., Chatterjee, A., Nussbaumer, A., Renzel, D., Scheffel, M., Friedrich, M., Santos, J. L., Duval, E., and Law, E. L.-C. (2011). Towards responsive open learning environments: the ROLE interoperability framework. In *Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning*, EC-TEL'11, pages 125–138, Berlin, Heidelberg. Springer-Verlag.

Hardt, D. (2012). The OAuth 2.0 Authorization Framework. RFC 6749, RFC Editor.

Saltzer, J. and Schroeder, M. (1975). The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308.

[12] http://www.imsglobal.org/toolsinteroperability2.cfm

Team, A. S. (2013). Apache shiro reference documentation. http://shiro.apache.org/reference.html.

Wilson, S., Sharples, P., Griffiths, D., and Popat, K. (2011). Augmenting the vle using widget technologies. *Int. J. Technol. Enhanc. Learn.*, 3(1):4–20.

Zibuschka, J., Herbert, M., and Roßnagel, H. (2010). Towards privacy-enhancing identity management in mashup-providing platforms. In *Proceedings of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy*, DBSec'10, pages 273–286, Berlin, Heidelberg. Springer-Verlag.

Annotated Trees and their Applications to XML Compression

Tomasz Müldner[1], Jan Krzysztof Miziołek[2] and Tyler Corbin[1]
[1]*Jodrey School of Computer Science, Acadia University, Wolfville, B4P 2A9 NS, Canada*
[2]*Faculty of Artes Liberales, University of Warsaw, Warsaw, Poland*
tomasz.muldner@acadiau.ca, jkm@al.uw.edu.pl, 094568c@acadiau.ca

Keywords: XML, Tree Compression using Annotated Trees, Permutation-based XML Compression.

Abstract: Permutation based XML-conscious compressors permute the input document to improve the compression ratio and support efficiency of operations, such as queries or updates. One such compressor, XSAQCT, uses the properties of the permuted document, called an annotated tree, to these operations. This paper provides the formal background for the definition of an of D. It also provides an algorithm for creating an annotated tree for the XML document and its reverse algorithm, and discusses a measure of compressibility using an annotated tree. The theoretical and algorithm approaches are followed by the experimental results showing compressibility of annotated trees and a general analysis of semi-structured data and XML compression.

1 INTRODUCTION

A **tree** is one of the most important and popular structures in computing, used to represent the relations between nodes. Therefore, there has been considerable research on succinct representation of trees while allowing various operations on these trees to be efficiently performed, see (Chen and Reif, 1996), (Bille et al., 2013), (Jacobson, 1989) and (Benoit et al., 1999). The eXtensible Markup Language, XML (XML, 2013), is one of the most popular data formats for the serialization of tree data structures and for the storage of relational data. Since XML documents are hierarchal and acyclic in nature, there have been numerous attempts to apply techniques used for general tree compression to XML, see e.g., (Busatto et al., 2005), (Ferragina et al., 2009) and (Busatto et al., 2008). A specific subset of tree-compressors designed specifically for XML (or semi-structured data in general) are called XML-conscious compressors, e.g., XQueC (Arion et al., 2007). These compressors typically parse the XML, and either sequentially build a model during compression or build a model and then compress the contents. XML-conscious *permutation-based* compressors permute the input document during a pre-order traversal and apply a partitioning strategy to group content nodes into a series of data containers compressed using a general-purpose compressor (a *back-end compressor)*. For example, XML compressor XSAQCT is a permuting and streamable XML compressor, see (Müldner et al., 2009). The compression starts with SAX-based parsing to permute the input into a compressed form of the XML tree, called an *annotated tree*. Then it is followed by storing data values in data containers and compressing data using a back-end compressors (such as GZIP (GZIP, 2013) or BZIP2 (bzip2, 2013)).

The annotated tree in XSAQCT can be considered to be a *high level index*, which was proved to be useful for various applications, e.g., updates, online rather than offline compression see (Corbin et al., 2013), and parallelization of the implementation. However, the formal definition of this mapping *was never provided* nor was the proof that it can be reversed. These questions are very important because without answering them, the compression process used by XSAQCT is not known to be *lossless*. This paper fills in these gaps, by providing a formal definition of a tree and an annotated tree, and the mapping τ from a labelled tree to the annotated tree, as well as the proof that τ is injective and can be inverted. An algorithm to create an annotated tree and its reverse are provided, followed by the discussion of the compressibility measure of this approach along with experimental results and a general analysis of XML compression.

Contributions. There are several novel contributions of this paper: (1) The theoretical part, i.e., the formal background for the definition of an annotated tree and a proof that the mapping from a tree to the

annotated tree is injective, and therefore, the annotated tree for the labelled tree D provides a faithful representation of D; (2) An application of of the annotated tree methodology with respect to XML, i.e., a formal proof that XSAQCT's compression process is lossless; (3) The algorithmic approach, i.e., the algorithm which inputs an arbitrary (cyclic or not) tree and outputs an annotated tree, and the "inverse" algorithm for XML, which inputs an annotated tree for the XML document, and outputs this document; (4) Quantification of text trees and a discussion of mutual information; (5) A compressibility measure defining the cost of using annotated trees and results of testing using an especially designed XML suite, showing high compressibility; and (6) General analysis of XML compression.

This paper is organized as follows. Section 2 introduces the formal background for this paper, including a definition of a tree and an annotated tree. Section 3 provides a description of trees with text elements. Section 4 provides algorithms implementing various mappings and shows time complexity of these algorithms. Section 5 provides results of testing and a general analysis, and finally Section 6 provides conclusions and describes future work.

2 TREES AND ANNOTATED TREES

2.1 Labeled Trees

Definition 1. *Let Σ be an alphabet, called the label alphabet. A **labeled tree** is an ordered tree with nodes labeled by strings from Σ^* and having arbitrary degrees (i.e., number of children).*

In what follows, by a *tree* we mean a labeled tree, and by **Trees** we denote the set of all trees. We use the letter D to denote a tree, with nodes denoted by lower-case letters x, y, u, v (with indices when needed) and by $x(a)$ we denote a node x labeled with a. Let $\mathbf{Label}(\mathbf{x})$ denote the label of the node x, $\mathbf{Height}(\mathbf{D})$ denote the height of D, $\mathbf{Nodes}(\mathbf{D})$ denote the set of all nodes of D, and $\mathbf{D_i}$ be the tree consisting of all nodes and edges in D at levels $1, \ldots, i, where 1 \leq i \leq height(D)$. Example of a tree is shown in Figure 1.

Definition 2. *A **path** in a tree is defined to be of the form $/x_1/x_2, \ldots, /x_k$, where x_1 is the root of D and for $1 \leq i < k, x_{i+1}$ is a child of x_i.*

We use a lower-case letter p (with indices when needed) to denote a path and $\mathbf{Paths}(\mathbf{D})$ to denote the set of all paths in D. For the path $p = /x_1/x_2, \ldots, /x_k$, let $\mathbf{Length}(\mathbf{p})$ be the number of nodes in the path p,

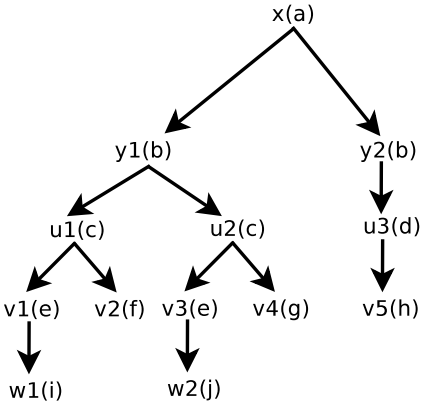


Figure 1: Example of a tree D.

$\mathbf{Last}(\mathbf{p})$ be the last element x_k, $\mathbf{Label}(\mathbf{p}) = Label(x_k)$ be the label of p, and for $Length(p) > 1$, let $\mathbf{p}\downarrow$ be the path p except the last element.

Definition 3. *Two paths in the tree $D, p_1 = /x_1/x_2 \ldots /x_k$ and $p_2 = /y_1/y_2 \ldots /y_n$ are called **similar** if $k = n$, $x_1 = y_1$ is the root of D and for $1 \leq i \leq k, Label(x_i) = Label(y_i)$.*

The similarity relation is an equivalence relation and we denote by $[\![\mathbf{p}]\!]$ the equivalence class of p, by $\mathbf{Similar}(\mathbf{D})$ the quotient set of this relation, i.e., the set of all different sets of similar paths. For $p \in Paths(D)$ let $\mathbf{Length}([\![\mathbf{p}]\!]) = Length(p)$ be the length of the equivalence class, $\mathbf{Label}([\![\mathbf{p}]\!]) = Label(p)$ be the label of the class (it is easy to see that both definitions of the length and the label of an equivalence class are well defined, i.e., independent of the choice of the path p from the equivalence class). Finally, for $q \in Similar(D)$ let $|\mathbf{q}|$ be the number of paths in q, and $\mathbf{Last}(\mathbf{q})$ be the sequence of nodes that are last elements of all paths in this class, *ordered* from left to right. Clearly, $|Last(q)| = |q|$.

In Figure 1, $Similar(D) = \{[\![/x]\!], [\![/x/y_1]\!], [\![/x/y_1/u_1]\!], [\![/x/y_1/u_3]\!], [\![/x/y_1/u_3/v_5]\!], [\![/x/y_1/u_1/v_1]\!], [\![/x/y_1/u_1/v_2]\!], [\![/x/y_1/u_2/v_4]\!], [\![/x/y_1/u_1/v_1/w_1]\!], [\![/x/y_1/u_1/v_3/w_2]\!]\}$ and for $p = /x/y_1, [\![p]\!] = \{/x/y_1, /x/y_2\}$; $Length([\![p]\!]) = 2$, $Label(p) = b$, $Last([\![p]\!]) = <y_1, y_2>$, and $|[\![p]\!]| = 2$.

Definition 4. *A partial relation $\prec$ in the set $Similar(D)$ is defined as follows: $[\![p_1]\!] \prec [\![p_2]\!]$ iff $Length([\![p_1]\!]) = Length[\![p_2]\!]) > 1, [\![p_1\downarrow]\!] = [\![p_2\downarrow]\!], Label([\![p_1]\!])$ is different from $Label([\![p_2]\!])$, and there exist paths $p1 \in [\![p_1]\!]$ and $p2 \in [\![p_2]\!]$ such that the node $Last(p1)$ is the left sibling of the node $Last(p2)$. The tree D has a **cycle** if there exist $q_1, q_2 \in Similar(D)$ such that $q_1 \prec q_2$ and $q_2 \prec q_1$. D is **acyclic** if it does not have a cycle.*

For D in Figure 1, $[\![/x/y_1/u_1/v_1]\!] \prec [\![/x/y_1/u_1/v_2]\!]$, while $[\![/x/y_1/u_1]\!]$ and $[\![/x/y_1/u_3]\!]$ are

not in relation $\prec$. There would be a cycle in D if there was a node $u_0(e)$ between nodes u_1 and u_2. By **Acyclic** we denote the set of all acyclic trees.

When it does not lead to confusion, a tree can be represented using a *simplified notation* (used e.g. for XML), by omitting node names, e.g., replacing $y_1(b)$ by b_1. In this notation, an equivalence class $[\![p]\!]$ will be denoted using the label $Label([\![p]\!])$, in upper case, e.g., for D from Figure 1, $[\![/x/y_1/u_1/v_1]\!]$ will be denoted by E.

2.2 Annotated Trees and g-trees

A tree D can permuted to create an annotated tree with nodes represented by equivalence classes of the similarity relation. In the worst case, if for each $q \in Similar(D)$, $|q| = 1$ then the size of D would be the same as the size of the corresponding annotated tree. However, typically XML documents are regular, i.e., for majority of paths $q \in Similar(D)$, $|q| \gg 1$ and the annotated tree provides a compressed representation of D. For a single tree D, there may be *more than one annotated tree* such that *each such annotated tree can be mapped back* to D. To formalize this idea, in this section we define annotated trees and annotated g-trees. Then we define two mappings, an injective mapping from the set of trees to the set of annotated g-trees and an injective mapping from the set of annotated g-trees to set of subsets of annotated trees. Finally, we define annotated text trees and a mapping from text trees to the annotated text trees. In this section we consider only acyclic trees, but in Section 4 we provide algorithms for all types of trees. In what follows, by a *dag* we mean an acyclic digraph.

Definition 5. *An **annotated tree** is an ordered tree with nodes additionally labeled by annotations (sequences of non-negative integers). An **annotated g-tree** is an unordered tree A (i.e., children are not ordered) such that (1) nodes of A are dags; (2) each dag $G \in Nodes(A)$ except the root has its nodes annotated; and (3) for every node $H \in Nodes(A)$ and for every child G of H there exists exactly one node $n \in Nodes(H)$, called the **source of G**, and different children of H have different sources.*

Nodes in the annotated tree are denoted by uppercase letters (with indices where appropriate), e.g., $\mathbf{X}(\mathbf{n})[\alpha_\mathbf{1}, \ldots, \alpha_\mathbf{j}]$ denotes a node X labeled with the label n and annotation $[\alpha_1, \ldots, \alpha_j]$ or in a simplified notation it is denoted as the node $N[\alpha_1, \ldots, \alpha_j]$. Let $\mathbf{AnnotationSum}(\mathbf{X}) = \sum_{i=1}^{j} \alpha_i$ be the sum of all annotations of the node X. By **Annotated** we denote the set of all annotated trees. Two examples of annotated trees (using a simplified notation) are shown in Figure 2 and Figure 3. As we will explain it later,

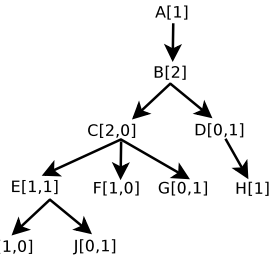


Figure 2: Annotated tree.

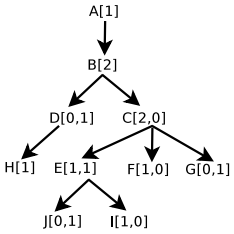


Figure 3: Another annotated tree.

both these trees represent the same tree D from Figure 1. Example of an annotated g-tree is given in Figure 4 (the source of a node is shown using a dashed arrow). By **Annotated−G** we denote the set of all annotated g-trees and by a *chain* we mean a rooted dag such that each node except the root has the in-degree one and each node except one designated node called the sink has the out-degree one; the sink has the out-degree zero. The reason for defining g-trees is that a dag G, which is a chain, will have its nodes representing children of the source (in left-to-right order) of G. If a dag G is not a chain then G needs to be topologically sorted for our usage. For example, in Figure 4 the topological sorting of the dag containing nodes E, F and G may produce the chain E, F, G and these three nodes can be made children of the source of this graph, i.e., the node C.

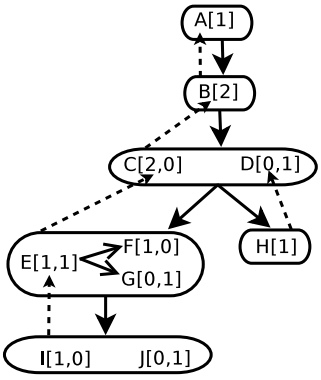


Figure 4: Example of a g-tree.

2.3 Tree Isomorphism

Since we will show that the mapping from the set of trees to the set of g-trees is injective, we need to define the concept of "identical" or isomorphic trees, which differ only in names of the corresponding nodes. We use a similar concept for annotated trees and g-trees.

Definition 6. *Two trees D_1 and D_2 are **isomorphic** iff they have the same height h and for each level $i, 1 \leq i \leq h$, the sequence of all nodes $<n_1, \ldots, n_j>$ (in left-to-right order) in D_1 at level i and the sequence of all*

nodes $<m_1,\ldots,m_k>$ (in left-to-right order) in D_2 at level $i, k = j$, and for $1 \le r \le k$, nodes n_r and m_r have the same degree and label.

2.4 Mapping Trees to Sets of Annotated Trees

We define the mapping $\tau : Acyclic \rightarrow Annotated{-}G$ in two steps; first for nodes and the tree structure, and then for annotations of nodes. Nodes in all dags are represented by equivalence classes of the similarity relation and written as $[\![p]\!](Label(p))[\alpha_1,\ldots,\alpha_j]$ or using a simplified notation $Label(p)[\alpha_1,\ldots,\alpha_j]$. If x is a node in D, which is not a root, then by $\mathbf{p_x}$ we denote a (unique) path $p \in Paths(D)$ which ends in x.

Definition 7. *Definition of mapping τ: $Acyclic \rightarrow Annotated{-}G$. Let $D \in Acyclic$.*

- *Mapping labeled nodes and the tree structure*
  1. *The root r of D is mapped to the root of $\tau(D)$ defined as a graph consisting of a single node $[\![/r]\!](Label(r))$.*
  2. *For any level $i > 1$ of D and equivalence class $q \in Similar(D)$ of length $i-1$ let $N_{q,i}$ denote the set of nodes x in D at level i such that $[\![p_x\downarrow]\!] \in q$. Clearly, the sets $\{N_{q,i} : q \in Similar(D)\}$ form a disjoint coverage of the set of all nodes in D at level i. Each set $N_{q,i}$ is mapped by τ to the single graph $G \in Nodes(\tau(D))$ at the level i; $G = \{[\![p_x]\!](Label(p_x)\}$ where the source of G is the node q. For any graph $G \in Nodes(\tau(D))$ and two nodes $q_1, q_2 \in G$, there is an edge $q_1 \Longrightarrow q_2$ in G iff $q_1 \prec q_2$, (see Definition 4). Since D is assumed to have no cycles, G is acyclic. The node $H \in Nodes(\tau(D))$ is the parent of the node G if the source of G belongs to the dag $Nodes(H)$. From the definition of sets $\tau(N_{q,i}), H$ is the unique parent of G.*
- *Annotations are defined by induction on the height of $\tau(D)$*
  1. *The annotation of the root is [1]*
  2. *Assume that annotations are defined up to the level $i, 1 \le i < Height(D)$ and for any equivalence class q of length i, $|Last(q)| = AnnotationSum(q)$, i.e., the sum of all annotations for the node q is equal to the number of last elements in all paths in this class. Consider a dag G at the level $i+1$ with the source X, and a node $Y \in Nodes(G)$. Let $Y = [\![p]\!](r)[\alpha_1,\ldots,\alpha_j], X = [\![p\downarrow]\!](m)$ and $k = AnnotationSum(X)$. First, we set the number j of annotations in Y to be k. From the inductive assumption it follows that the sequence $Last([\![p\downarrow]\!])$ has k elements $s_1,\ldots,s_k$. For $1 \le j \le k$, we define α_j to be the number of children in D of the node s_j which are labeled with r. It is easy to see that $|Last([\![p]\!])| = AnnotationSum(Y)$.*

Let $\mathbf{Trees}_\tau$ denote the image $\tau(Acyclic) \subset Annotated{-}G$. From the Definition 7, it follows that *any* g-tree $A \in Trees_\tau$, $A = \tau(D)$ has the following properties:

1. $Height(D) = Height(A)$
2. For $1 \le i \le Height(D)$, D and A have identical sets of labels at level i
3. For a dag $G \in Nodes(A)$, let X be the source of $G, X = [\![q]\!]$ and $k = AnnotationSum(Y)$. Then for each node $Y \in Nodes(G)$, $Y = [\![p]\!](n)[\alpha_1,\ldots,\alpha_j]$ there exists $i, 1 \le i \le k$ such that $\alpha_i > 0$; $j = k$, and $[\![q\downarrow]\!] = [\![p]\!]$
4. If a node q is not a source of any dag, then all nodes in the sequence $Last(q)$ are leaves in the tree D.

For the tree D from Figure 1, the g-tree $\tau(D)$ is shown in Figure 4. From Property 1 it follows that the height of a g-tree for the document D is the same as that of D; however typically these trees have different width and therefore represent a compressed form (see more in Section 5).

Theorem 1. *The mapping $\tau : Acyclic \rightarrow Trees_\tau$ is injective.*

Proof. We will define the mapping $\tau^{-1} : Trees_\tau \rightarrow Acyclic$ such that $\tau^{-1} \circ \tau = \tau \circ \tau^{-1}$ is an identity mapping, i.e., it maps a tree to an isomorphic tree. Let $A \in Trees_\tau$. Then τ^{-1} is defined inductively on levels i of A:

1. for $i = 1$, let the root of A be a graph G consisting of a single node $X(a)[1])$. Then $\tau^{-1}(G) = x(a)$
2. Assume that for all levels $i, 1 < i < Height(A), D_i = \tau^{-1}(A)_i \in Acyclic$. We define $\tau^{-1} : A_{i+1} \rightarrow D_{i+1}$. Let G be a dag in A at level $i+1$, and $Nodes(G) = \{Y_1,\cdots,Y_m\}$, where for $j, 1 \le j \le m$, $Y_j = [\![q_j]\!](r_j)[\alpha_1^j,\ldots,\alpha_k^j]$, the node $X = [\![p]\!](m)$ be the source of of G, $k = AnnotationSum(X)$, and $Last([\![p]\!])$ be the tuple $s_1,\ldots,s_k$.
   For $h, 1 \le h \le k$ we now define $\tau^{-1}(G)$ as nodes in D at the level $i+1$ that are children of each node s_h. First, for $j, 1 \le j \le m$, let $N_{h,j}$ be the set of nodes in D at the level $i+1$ consisting of α_h^j nodes in D labeled by r_j. We define a partial order between these sets as follows: for j_1 and $j_2, 1 \le j_1, j_2 \le m$, $N_{h,j_1} \prec N_{h,j_1}$ iff in G there is

an arc between $[\![q_{j_1}]\!]$ and $[\![q_{j_2}]\!]$. Finally, children of s_h are nodes from sets $N_{h,1},\ldots,N_{h,m}$ defined as follows: (1) within each set $N_{h,j}$, nodes are arbitrarily ordered and appear one after another; (2) for any two sets N_{h,j_1} and N_{h,j_2} all nodes from N_{h,j_1} appear to the left of all nodes in N_{h,j_2} iff $N_{h,j_1} \prec N_{h,j_2}$; and (3) for any set $N_{h,j}$ which is not in the relation $\prec$, we arbitrarily place children from this set to appear after all nodes from sets which are in this relation with another set.

Clearly the tree $\tau^{-1}(A)$ is acyclic and for $D \in Acyclic, \tau^{-1}\circ\tau(D) = D$, for $A \in Trees_\tau$, $\tau^{-1}\circ\tau(A) = A$. □

Next, we define the mapping $\gamma : Trees_\tau \to 2^{Annotated-G}$. Let $TS(G)$ be the set of all topological sortings of a dag G, and for $G_1,\cdots,G_n$ $P(G_1,\ldots,G_n)$ be a Cartesian product $\times_{i=1}^{n} TS(G_i)$. If $<g_1,\ldots,g_n> \in P(G_1,\ldots,G_n)$ then each graph g_i represents a topologically sorted graph G_i.

Definition 8. *Mapping γ is defined as follows: For $T \in Trees_\tau$, $\gamma(T) = \{\gamma_{<g_1,\ldots,g_n>}(T) : <g_1,\ldots,g_n> \in P(G_1,\ldots,G_n)\}$ where $\gamma_{<g_1,\ldots,g_n>}(T)$ is the g-tree T with all graphs $G_1,\ldots,G_n$ replaced respectively by graphs $g_1,\ldots,g_n$, and having the same arcs between dags (as well as sources) as in the tree T.*

Let $\mathbf{Trees}_\gamma$ denote the image $\gamma(Trees_\tau) \subset 2^{Annotated-G}$.

Theorem 2. *The mapping $\gamma : Trees_\tau \to Trees_\gamma$ is injective.*

Proof. We will define the mapping $\gamma^{-1} : Trees_\gamma \to Trees_\tau$ s.t. $\gamma^{-1}\circ\gamma$ and $\gamma\circ\gamma^{-1}$ are identity mappings. It is sufficient to show that each topologically sorted dag g_i with annotated nodes can be uniquely mapped to the dag G_i. Two different nodes $X[\alpha_1,\ldots,\alpha_k]$ and $Y[\beta_1,\ldots,\beta_k]$ from the dag g_i will be called *dependant* if there exists $m, 1 \le m \le k$ such that $\alpha_m > 0$ and $\beta_m > 0$. Now, let us define $\gamma^{-1}(g_i)$ to be the graph consisting of the same nodes as in the graph g_i and with the same source, but with arcs defined as follows: there is an arc $X \Longrightarrow Y$ iff the node X appears in the topological sort used in g_i before the node Y, and X and Y are dependant. □

It is easy to see that each g-tree $T \in Trees_\gamma$ is in one-to-one correspondence with an annotated tree. Since $\tau : Acyclic \leftrightarrow Trees_\tau$ and $\gamma : Trees_\tau \leftrightarrow Trees_\gamma$, each tree D can be mapped to the set of annotated trees, denoted by $\mathbf{Annotated(D)}$ and defined by the composition of τ and γ. It is easy to see that every annotated tree from the set $Annotated(D)$ represents

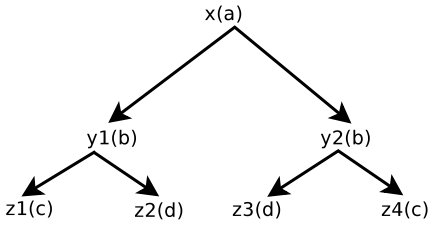


Figure 5: Cyclic Tree.

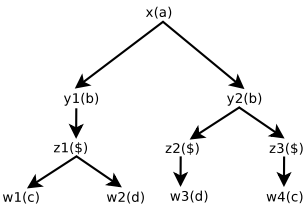


Figure 6: Acyclic tree with dummy nodes.

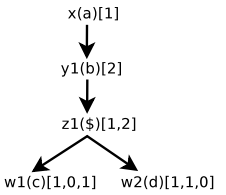


Figure 7: The annotated tree with dummy nodes.

D. For example, both annotated trees shown in Figure 2 and Figure 3 belong to the set $Annotated(D)$ that uniquely represents the tree D.

If there is a cycle in D, then we map D to an acyclic tree with the so-called dummy nodes, denoted by \$. After adding a dummy node to a cyclic document D in Figure 5, this tree will be acyclic, see Figure 6, and it can be mapped to the annotated tree shown in Figure 7. We do not formally prove that the mapping from the tree with cycles to a tree with the dummy nodes is injective but provide Algorithm 1 in Section 4 showing how cycles can be removed.

3 TEXT TREES AND THEIR COMPRESSION

Since the procedure of compressing text trees is almost identical to the procedure of compressing labeled trees, in this section we provide only the description of how text nodes are dealt with.

3.1 Text Trees

Definition 9. *Let Δ be an alphabet, called the text alphabet, its elements are $\backslash 0$ terminated strings. A **text tree** is a tree with two kinds of nodes; element nodes labeled by strings from Σ^* (see Definition 1) and text nodes labeled by strings from Δ^* such that text nodes are always leaves, the root of the text tree is an element node, and any two sibling text-nodes are separated by at least one element node.*

For text trees we use the same notations as for trees; text labels are denoted using letter t (with in-

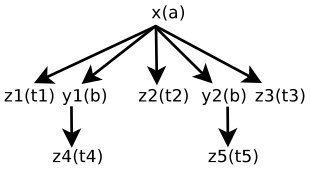


Figure 8: Text tree.

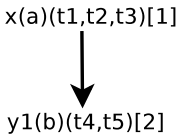


Figure 9: The annotated text tree.

dices if needed), see Figure 8. By **TextTrees** we denote the set of all text trees, and by **AcyclicTextTrees** we denote the set of all acyclic text trees. A text tree can be used to represent an XML document with text values represented by labels of text nodes. Text nodes may or may not be present and now we define complete text trees corresponding to the full-mixed content of XML documents, see (Müldner et al., 2012).

Definition 10. *An element leaf node in the text tree is the element node that has no element child. A text tree is called* ***complete*** *if every non-leaf element node has the left and the right text sibling, and every element leaf node has exactly one text child node.*

The text tree from Figure 8 is complete; it would not be complete if any of the text nodes were missing. Note that in XSAQCT when the input XML document D is parsed then for any missing text node, a text node labeled by an empty text (consisting of $\backslash 0$ only) is added. To support a unique representation of an XML document using this technique, added text nodes are removed while D is restored. In what follows, we assume that *text trees are complete*.

3.2 Compressed Representation of Text Trees

Definition 11. *For a node X in the annotated tree with m children $Y_1,\ldots,Y_m$, $m \geq 0$ let* **Number**$(\mathbf{X})$ $= (\sum_{j=1}^{m} AnnotationSum(Y_j)) + AnnotationSum(X)$. *An* ***annotated text tree*** *is an annotated tree with nodes additionally labeled by concatenations of strings from Δ^*, called text labels, such that a text label T of a node $X(a)(T)$ is equal to the concatenation of $Number(X)$ text labels.*

Example of an annotated text tree is shown in Figure 9. By **AcyclicAnnTextTrees** we denote the set of all acyclic annotated text trees.

3.3 Mapping Text Elements

For every equivalence class $[\![p]\!]$, labels of text nodes that are children of element nodes (in left to right order) from the sequence $Last([\![p]\!])$ will be concatenated into a single text label of the node in the annotated tree that corresponds to this equivalence class. For example, the text tree shown in Figure 8 will be mapped to the annotated text tree from Figure 9, where t_1, t_2 denotes a concatenation of texts. The reason for mapping text nodes this way is that for querying of the XML documents, the query of the form $/X$ returns the concatenated texts appearing in this path.

Definition 12. *Let $D \in AcyclicTextTrees$ and let A be the image of D under τ (see Definition 7) as if there was no text nodes in D. Now, for every node $X \in A$, we will define its text label T such that the number of texts concatenated in T will be equal to $Number(X)$. Let $\tau^{-1}(X)$, as defined the proof of Theorem 1, be the sequence $x_1,\ldots,x_k$ of element nodes, where $k = AnnotationSum(X)$ and let us consider two cases:*

1. *X is a leaf. Then for every $1 \leq i \leq k$, x_i has exactly one text child, and let T be the concatenation of labels of these k children. Clearly, $Number(X) = AnnotationSum(X) = k$.*
2. *X is a not leaf, and it has m children, $Y_1,\ldots,Y_m$. Thus, there are $Number(X)$ text children of nodes $x_1,\ldots,x_k$, and we define T to be the concatenation of labels of these children (in left-to-right order).*

Theorem 3. *The mapping of text elements is injective.*

Proof. Let A be an annotated text tree and D be the text tree; we will describe how text labels from A are mapped into the text elements in D. Consider the node $X \in Nodes(A)$ and let $k = Number(X)$. The image of X under the reverse mapping is the sequence of element nodes $x_1,...,x_k$. We consider two cases:

1. $X(T)$ is a leaf. Then $k = Number(X) = AnnotationSum(X)$ and T is a concatenation of k texts $t_1,\ldots,t_k$. For every $i, 1 \leq i \leq k$, we define a text child $y_i(t_i)$ of the node x_i.
2. $X(T)$ is a not leaf and it has m children $Y_1(T_1),\ldots,Y_m(T_m)$, where T is a concatenation of $r = (\sum_{j=1}^{m} AnnotationSum(Y_j)) + AnnotationSum(X)$ texts $s_1,\ldots,s_r$. Given a node x in a tree from $AcyclicTextTrees$ with k element children $y_1,\ldots,y_k$ and $k+1$ texts $t_0,\ldots,t_k$, text nodes are defined as follows: (1) the node labeled t_0 is the leftmost child of the node x, and (2) for $i, 1 \leq i \leq k$, the node labeled t_i is the right sibling of the node y_i. Assume that each element node x_i has n_{r_i} element children, so $r = (\sum_{j=1}^{k} n_{r_j}) + k$. Now, let us create text children of nodes $x_1,\ldots,x_k$ using the above method and consecutive groups of texts from T, i.e., for x_1 and its element children $y_1,\ldots,y_{n_{r_1}}$ we use the first $n_{r_1}+1$ texts from T,

for x_2 and its element children $y_2,\ldots,y_{n_{r_2}}$ we use the next $n_{r_2}+1$ texts from T, etc.

Clearly the tree $\tau^{-1}(A)$ is acyclic and for $D \in Acyclic(D), \tau^{-1}\circ\tau(D) = D$. □

4 ALGORITHMIC APPROACH

This section provides the algorithm to create an annotated tree for a fixed labeled tree D, which may or may not be cyclic. There are two steps in the process of creating an annotated tree, respectively implemented by Algorithm 1 and Algorithm 2. After the first step, when parsing of D is performed, for each set of similar paths there will be an associated graph of annotated nodes, which may include (in case of a *cyclic* input document) a special annotated *dummy symbol* \$. The second step uses data created by the first step to create an annotated tree. We also show the Algorithm 3 for the inverse mapping for XML, which inputs an annotated tree for the XML document D and outputs D.

4.1 Notations and Auxiliary Abstract Data Types

Recall that $x(n)$ denotes a node x labeled by n in D, and $X(n)[A]$ denotes an annotated node X labeled by n with the annotation A. For the path $p \in P$ and a node $x \in D$, let $\mathbf{p/x}$ denotes the path p extended by $/x$, and $\mathbf{X0(p)}$ denotes the so-called *current node* (which could be empty). Recall from Section 2.1 that $p\downarrow$ denotes the path p except its last element, and from Section 2.2 that $AnnotationSum(X)$ denotes the sum of all annotations of the node X. Let $\mathbf{A(n)}$ denote a list of $AnnotationSum(X0(p\downarrow))-1$ occurrences of integer n, i.e., $k-1$ occurrences of n, where k is the sum of all annotations of the current node for the path $p\downarrow$. For example, if $AnnotationSum(X0(p\downarrow))-1$ is equal to four, then $[A(5),1]$ denotes the annotation [5,5,5,5,1]. If X is a node or the dummy node then by $\mathbf{Inc(X,i)}$ we denote the following modification of the annotation of X: if $i=1$ then the last annotation of X is incremented by one; if $i=0$ then ",0" is added after the last annotation of the dummy symbol. Let $\mathcal{P}$ be the set of equivalence classes of the similarity relation (see Definition 3) maintained by the Algorithm 1 (initially, it consists of the equivalence class for the path /root of D). For every equivalence class $q \in \mathcal{P}$, let $G(q)$ be a pair consisting of (a graph with annotated nodes, a dummy node \$), and let \$(p) denotes the annotation of the dummy node for the class $[\![p]\!]$. Now, we define several auxiliary notations and operations: (1) `int insertP(path p)` returns 0 if $[\![p]\!] \in \mathcal{P}$; otherwise it inserts $[\![p]\!]$ into $\mathcal{P}$, sets $G([\![p]\!])$ to be an empty graph and empty dummy symbol, and returns 1; (2) `Annotatednode insertG(path p, label n, annotation A)` ; (3) `node memberG(path p, label n)` returns the unique node X in $G([\![p]\!])$ such that X is labeled with n; (4) `addArc(node X1, node X2, path p)` adds a new arc connecting $X1$ and $X2$ in $G([\![p]\!])$; (5) `annotation Ann(path p)` if the path p is of the form \root or S(X0(p↓)) is equal to 1 then return 1, otherwise return the annotation S(X0(p↓))-1 of 0's, 1; (6) `int reachableG(node X1, node X2, path p)` returns 1 if the node $X2$ is reachable from the node $X2$ in $G([\![p]\!])$, otherwise it returns 0; (7) `sortG(path p)` performs a topological sort of $G([\![p]\!])$; (8) `update(path p)` for each node X in $G([\![p]\!])$ performs $Inc(X,0)$.

4.2 Mapping a Labeled Tree D to an Annotated Tree

This section presents Algorithms 1 and 2 which map a labeled tree D, possibly cyclic, into an annotated tree, which in case of cycles adds dummy nodes labeled by \$ (thus algorithms in this section are designed for *arbitrary trees*, rather than for XML trees). Algorithm 1 performs a depth-first search traversal of the input tree, moving down and up. For the XML document D, these actions would be triggered by entering the beginning of the element, i.e., `<x` and the end of the element, i.e., `</x`. The algorithm maintains the current path $p0$ in D, and for each path p in D it maintains the set $\mathcal{P}$ of paths and associated graphs $G([\![p]\!])$. It also maintains the current annotated node $X0(p)$ in $G([\![p]\!])$ and the annotated dummy node.

The Algorithm 1 is initialized by setting the node $x0$ to be the root of D and $p0$ to be the path $/x0$. Then it performs a loop moving down and up until it reaches the root node and there *are no more un-visited children of the root*, while maintaing the following two invariants: (1) The current node $X0(p0\downarrow)$ is not null; (2) If the algorithm moves down to the node x, such that the path p/x already exists, then there exists a unique annotated node X in the graph $G([\![p0]\!])$ which has the same label as the label of x. Once the algorithms complete their actions, the set $\mathcal{P}$ stores all equivalence classes of the similarity relation for D.

Table 1 shows the trace of the execution of Algorithm 1 for the document D from Figure 5, using simplified notation. The current annotated node X0(p) (if any) is underlined. If the tuple (path q, G(q)) has not changed, then it is not shown again. The action of going down to the node x is shown as ↓x and the action of going up to the node x is shown as ↑x. Each $q \in \mathcal{P}$

will represent a node of the annotated tree, and the annotated nodes from the graph $G(q)$ will represent children of these nodes. Algorithm 2 inputs the output of the Algorithms 1 and produces the annotated tree.

Algorithm 1: Algorithm which maps a tree to a set $\mathcal{P}$ and associated graphs.

Require: x0 = root of D, p0 = /x0, G($[\![p0]\!]$)=(graph and dummy node; both empty)

```
 1: function TRAVERSE
 2: | while true do
 3: | | if current is root and no more nodes then
 4: | | | return;
 5: | | end if
 6: | | if moving down to x then
 7: | | | p1 = p0/x;
 8: | | | if insertP(p1) then // p1 is a new path
 9: | | | | X1 = insertG(p0, label of x, [A(0),1]);
10: | | | | if X0(p0) ≠ ∅ then
11: | | | | | addArc(X0(p0), X1, p0);
12: | | | | end if
13: | | | else// p1 is already defined
14: | | | | X1 = memberG(p0,label of x);
15: | | | | if X0(p0) ≠ ∅ AND X1 ≠ X0(p0) then
16: | | | | | // check if arc can be added
17: | | | | | if !reachableG(X0(p), X1, p0) then
18: | | | | | | // can add
19: | | | | | | Inc(X1,1);
20: | | | | | | update(p1);
21: | | | | | | addArc(X0(p0), X1, p0);
22: | | | | | | if $(p0) ≠ ∅ then
23: | | | | | | | Inc($ ,0);
24: | | | | | | end if
25: | | | | | else// needs a dummy node
26: | | | | | | if $(p0) ≠ ∅ then
27: | | | | | | | Inc($,1);
28: | | | | | | else
29: | | | | | | | $(p0) = $[A(1), 2];
30: | | | | | | end if
31: | | | | | | update(p0);
32: | | | | | | Inc(X1,1);
33: | | | | | end if
34: | | | | else
35: | | | | | X0(p0) = X1;
36: | | | | | Inc(X1,1);
37: | | | | | update(p1);
38: | | | | | if $(p1) ≠ ∅ then
39: | | | | | | Inc($ ,0); Inc($ ,1);
40: | | | | | end if
41: | | | | end if
42: | | | end if
43: | | | X0(p0) = X1;
44: | | | p0 = p1;
45: | | else// moving up to the node x
46: | | | X0(p0) = ∅;
47: | | | p0 = p0↓;
48: | | end if
49: | end while
50: end function
```

Let n be the number of nodes in the input tree. The Algorithm 1 is based on DFS-traversal of a tree, which has $O(n)$ time complexity, with the nested call (line 17) to the function *reachableG*(), which has $O(|V|+|E|)$ time complexity, where $|V|$ is the number of vertices in the graph passed as a parameter to the function, and $|E|$ is the number of vertices in this graph. Therefore, time complexity of this algorithm is $O(n^2)$. Similarly, Algorithm 2 has $O(n^2)$ complexity.

Table 1: Trace of the execution of Algorithm 1.

Move	p0	p1	$[\![p]\!]$: G($[\![p]\!]$), dummy
	/a		A: ∅
↓b1	/a/b1	/a/b1	A: B[1], ∅
↓c1	/a/b1/c1	/a/b1/c1	B: C[1], ∅
↑b1	/a/b	/a/b1/c1	
↓d1	/a/b1/d1	/a/b1/d1	B: C[1]→D[1], ∅
↑b1	/a/b1/		
↑a	/a		B:, C[1]→D[1], ∅
↓b2	/a/b2	/a/b2	A:Y(b)[2], ∅
			B: C[1,0]→D[1,0], ∅
↓d2	/a/b2/d2	/a/b2/d2	B: C[1,0]→D[1,1], ∅
↑b2	/a/b2		
↓c2	/a/b2/c2	/a/b2/c2	B: C[1,0]→D[1,1]
			B: C[1,0,1]→D[1,1,0]
			$[1,2]
			$[1,2]
↑b2	/a/b2		
↑a	/a		

Algorithm 2: Algorithm which maps a set $\mathcal{P}$ and associated graphs to an annotated tree.

Require: Initially p = \root, X = root(label of the root of D)[1]

```
 1: function FINALIZE(Path p, Node X)
 2: | if $(p) ≠ ∅ then
 3: | | make the node $(p) a child of X;
 4: | | X = $(p);
 5: | end if
 6: | Node R = sort(p); // now G(p) is a chain
 7: | make annotated nodes in G(p) children of X
 8: | for every node X1 in G(p) do
 9: | | if G(p) ≠ ∅ then
10: | | | Finalize(p/(label of X1), X1);
11: | | end if
12: | end for
13: end function
```

4.3 Mapping an Annotated Tree to the Labeled Tree

Now we present the algorithm 3, which describes for an XML document a mapping reverse to the previously described mapping, i.e., it inputs an annotated tree (possibly with dummy nodes) and outputs the XML document. The following notations are used: (1) ann(n): first digit in the annotation of the node n; (2) chop(n): remove the first digit in the annotation of n (always 0); (3) dec(n): decrement by one the first

Algorithm 3: Algorithm which inputs an annotated tree and outputs the XML document.

```
Require: Called for the root of the annotated tree
 1: function RESTORE(Node c)
 2: |  n = LC(c);
 3: |  while n ≠ ∅ do
 4: |  |  if ann(n) > 0 then
 5: |  |  |  if n is not a dummy node then
 6: |  |  |  |  output < + label of n + >;
 7: |  |  |  end if
 8: |  |  |  Restore(n);
 9: |  |  else
10: |  |  |  chop(n);
11: |  |  end if
12: |  |  n = RS(n);
13: |  end while
14: |  dec(c);
15: |  if c is not a dummy node then
16: |  |  output </ + label of c + >;
17: |  end if
18: |  if ann(c) = 0 then
19: |  |  chop(c);
20: |  else
21: |  |  output < + label of n + >;
22: |  |  Restore(c);
23: |  end if
24: end function
```

digit in the annotation of n (never 0); and (4) LC(n) and RS(n): respectively the leftmost child and right sibling of n. Given the annotated tree T, to restore the XML document, the following actions should be executed: `output <root label>; Restore(root of T); output </root label>` Algorithm 3 has $O(n^2)$ complexity.

5 EXPERIMENTAL RESULTS AND GENERAL ANALYSIS

This section describes an analysis of distribution of text elements in semi-structured data, and the definition of a compressibility measure for annotated followed by experimental results using an XML suite. Finally, it provides a general analysis of XML compression.

5.1 Quantification of Text Trees and Mutual Information

The use of annotated representation of a tree T implies the following hypothesis about the distribution of text elements in the tree: *Each equivalence class in a tree T also defines a unique random variable in which the text children are sampled from.* The hypothesis does not imply anything about the mutual information within the set of random variables; however, there are some implicit consequences as to how mutual information is dealt with. Specifically, *mutual information* is the amount of information that one random variable contains about another random variable; or the amount of reduction in uncertainty of one random variable due to the knowledge of the other. We define mutual information as $\mathbf{I}(\mathbf{X}|\mathbf{Y})$, the reduction in the uncertainty of X due to the knowledge of Y and consider two general cases in the transfer of information: (1) $[I(\text{Child}|\text{Parent}^N)]$ - the transfer of information from a parent node to each of its children, or the transfer of information from a Nth ancestor to each of its Nth descendants; (2) $[I(\text{Sibling}_i|\text{Sibling}_j)|\forall j < i]$ - the transfer of information from a sibling node to each of its prior siblings. Although we cannot explicitly prove any general relationships of the two clauses above, we can extract information about the general use of semi-structured data and their affects on these relationships.

$\mathbf{I}(\mathbf{Child}|\mathbf{Parent})$: With respect to most semi-structured data formats, e.g., XML and JSON, the text of non-leaf elements is mostly the whitespace data to make the semi-structured data human-readable, i.e., there is almost no relation between the information in any parent and the information of a leaf node. However, there is a clear relation between the information in any parent and the information of any non-leaf child. The data is only whitespace, and we can restrict the text alphabet to ASCII: SPC (0x20), TAB (0x09), LF (0x0a), VT (0x0b), FF (0x0c), and CR(0x0d). One caveat to this general statement, for example, are formatting tags in HTML, such as `"display <b>bold</b> text"` . However, in the optimal case, and for semantic equivalence of the XML, we can functionally, and not statistically, relate the whitespace characters and the depth of the node, i.e., the depth multiplied by an indent (tab, sequence of whitespace, etc.)

$\mathbf{I}(\mathbf{Sibling_i}|\mathbf{Sibling_j})$: The relationship among siblings is slightly more complicated. If we consider the set of text elements for each equivalence class of leaf nodes, we can describe similarity between two sets using Statistical and Alphabetical Similarities, Functional Relations, and Temporal/Semantic and Structural Relations. Although statistical and alphabetical similarities form the basis of our hypothesis about the relationship of data among individual equivalence classes, they can also be used describe the relationship of data across equivalence classes. For example, two nodes: `LastName` and `FirstName` would be highly informationally related. However, if two tags consist of only free-formed English, the alphabets

may be similar, but the words, sentences, etc. may be different. Therefore, while there may be some statistical relationship among the character frequency, it quickly declines as we increase the degree of our statistics. Functional relations would describe the tags whose information is just some deterministic function of another tag. For example, a `text` tag and a `sha-256` tag are functionally related. Temporal/Semantic and Structural Relations, while being a form of statistical similarity, describe the tags that have some temporal or sequential relation with its siblings. Consider this example: `<Questions>` `<Q>How many bits to an octet?</Q>` `<A>There are 8-bits to an octet.</A>` `<Q>How many bits to a byte?</Q>` `<A>Generally speaking, 8.</A>` `</Questions>` . Although functional relations can be exploited, they require the relationship to be known before compression, i.e., there is an underlying schema to the semi-structured data. Thus, these type of similarities are often not considered for general-purpose compression. With respect to non-leaf children, we expect the data to be quite statistically (and structurally) similar, because the data of a sibling set would be consistently formatted for human readability (or lack thereof). However, with leaf children, the annotative representation of a text tree will only exploit the sequence of character data local to an equivalence class and will not consider the sequence of character data local to some subtree (as shown above). Therefore, no compression algorithm will be able to use the knowledge of "How many bits to an octet" to compress the information of "There are 8-bits to an octet.". However, to compress "How many bits to a byte", we can use the knowledge of previously asking "How many bits to an octet". Therefore, semantically related text of the same vertex would have a high statistical similarity, whereas semantically related text of the same subtree will not be considered for compression.

5.2 Experimental Results

Throughout this section we use the following notations: D is a tree and A belongs to the set of annotated trees $Annotated(D)$. Recall from Section 2.2 that this set uniquely represents D, therefore the description provided in this section does not depend on the choice of A. For any tree T (annotated or not), by $|\mathbf{T}|$ we denote the number of nodes in T. Let the width of D at any level i be denoted by $\mathbf{width}(\mathbf{D},\mathbf{i})$ and let $\mathbf{Ann}(\mathbf{X})$ denote the annotation list of the node $X \in Nodes(A)$ and let $|\mathbf{Ann}(\mathbf{X})|$ denote the length of the annotated list. Finally let $X_1, X_2, \ldots, X_{N_i}$ be all nodes in A at level i (i.e., $width(A,i) = N_i$).

From the construction of an annotated tree, it follows that (see also Properties 2.4): (1) $width(D,i) = \sum_{k=1}^{N_i} AnnotationSum(X_k)$; and (2) In A, for any node X and its child Y, $AnnotationSum(X) = |Ann(Y)|$. Therefore, the larger the sum of all annotations of X, the longer the annotation list of Y. The implication of zeros appearing in the annotations of node X is two-fold: (a) in A, they shorten the length of the annotation list; (b) in D, Y does not appear as a child of X. If D was "completely regular" and there were no missing children, then there would be no 0s on the annotation lists. From our experiments, it follows that a leaf annotation lists are usually very long, while for the leaf's ascendants, annotations they are getting progressively shorter. To analyze this phenomenon consider a leaf node X in A, its parent Y, and Y's parent Z (grandparent of X). Since $AnnotationSum(Y) = |Ann(X)|$, the annotation list $Ann(X)$ must have been increased because of the structure of D, specifically Y has "very often" appeared as a child of Z, likely multiple times (resulting in annotations greater than 1). On the other hand, many 0s appearing in $Ann(X)$ indicates that X has "very rarely" appeared as a child of Y.

Let us now consider the **compressibility measure**, which measures the cost of storing the annotated tree A compared to the cost of storing the original document D. Our definition is to be implementation-independent and data are *not compressed* using a backend compressor. As discussed before, in general A has fewer nodes than D, but there is an additional cost of storing annotation lists. Let C be the cost of storing a single information, such as a single integer annotation or a node label. Therefore, the storage cost of *a single tree node* is $3 \times C$, and the cost of storing an annotation list of length L is $L \times C$. The total storage cost of D is equal to the cost of storing all nodes in A, including all annotation lists over the cost of storing all nodes in D, resulting in $\frac{(3\times C\times|A|)+(\sum_{X\in A}|Ann(X)|)\times C}{3\times C\times|D|}$. Performing a few simplifications, gives the following formula (independent of the cost C):

Definition 13. *The compressibility measure is defined as follows:*

$$\mu_c(D) = \frac{|A|+(\sum_{X\in A}|Ann(X)|)/3}{|D|}.$$

In this section we provide experimental results showing values of the measure from Definition 13 applied to a suite of XML files . Characteristics of these files are shown in the first four columns of Table 2, where V(N) denotes the value V*10^{N}, Size is the size of file in Bytes, E:A denotes the number of elements and attributes, AC denotes the number

Table 2: Overview of XML Test Suite and Results of Testing.

XML File	Size	E:A	AC	AT	C
1gig	1.17(9)	1.6(7):3.83(6)	2.05(8)	680	0.41
BaseBall	6.72(5)	2.8(5):0	6.6(5)	47	0.78
enw.books	1.56(8)	5.3(6):4.9(5)	6.38(5)	29	0.40
enw.latest	5.96(9)	1.84(9):1.85(8)	2.59(9)	39	0.47
lineitem	3.22(7)	1.02(6):1	9.6(6)	19	0.31
UniProt	1.15(8)	9.86(5):1.44(9)	1.05(9)	217	0.36

of annotations, AT denotes the number of nodes in the annotation tree, and C denotes the compressibility measure. The files are 1gig.xml (a randomly generated XML file, using xmlgen (xmlgen, 2013)), baseball.xml (Baseball.xml, 2013), enwikibooks.xml and lineitem.xml from the Wratislavia corpus (Corpus, 2013), and uniprot_sprot (Consortium, 2013). This suite has been chosen because XML files included there have an ability to represent specific extremes of semi-structured data. For example, enwiki-latest.xml, the current revision of English Wikipedia, while being a very large document, encompasses two extremes: the distribution of character data is very non-uniform (i.e., the majority of the data falls within one node) and that path is predominantly free-formed English. Conversely, uniprot_sprot.xml is a highly uniform XML file (i.e., the data is evenly distributed), and the file is predominantly markup. The file 1gig.xml has the property that the subtree entropy is extremely low (subtrees are quite similar); however, each subtree differs by a parent node (for example, `/a/b/d/e/f` vs. `/a/z/d/e/f`). The file lineitem.xml, has the property that it is an incredibly regular tree (few missing nodes), and in addition, has a nice mixture of text and numeric data. The file enwikibooks.xml is quite structurally similar to enwiki-latest.xml but is a fraction of its size. Finally, baseball.xml is an extremely irregular XML file. The last column of the Table 2 shows the test results, specifically values of the compressibility measures (see Definition 13) for the six XML file. These results show that the annotated tree provides a well-compressed representation of the original files, even in the presence of very large files.

5.2.1 Analysis of XML Compression

In measuring the compressibility of the annotated transform, each XML file was transferred into three separate files: (1) the annotated tree (a list of strings encoded in depth-first ordering); (2) the annotations (written in depth-first ordering); and (3) the text containers (written in depth-first ordering), see Algorithm 4. Consequently, this transform does nothing but to act as a pre-processor for other text compressors.

Ignoring Kolmogorov complexity (and Kolmogorov compressors), we assume the XML data to be distributed according to some random variable (or set of random variables). Therefore, the ideal way to analyze the compression benefits induced by the annotated transform would be to analyze how much benefit, with respect to the absolute lower bound, was obtained. However, since it is nearly impossible to calculate the exact entropy of the XML sources, the next ideal step would be to approximate the entropy using a lossless compression algorithm, which would be infeasible for the analysis of the proposed transform. Analyzing the percentage increase in compression does not take into consideration how substantial a percent decrease in size is, but it will be used as the base metric of comparison. For this analysis, a series of different compressors were used, see Figure10. LZ77-based (Ziv and Lempel, 2006) compressors, GZIP[1] and XZ[2], BWT-based (Burrows and Wheeler, 1994), BZIP2[3], Predic-

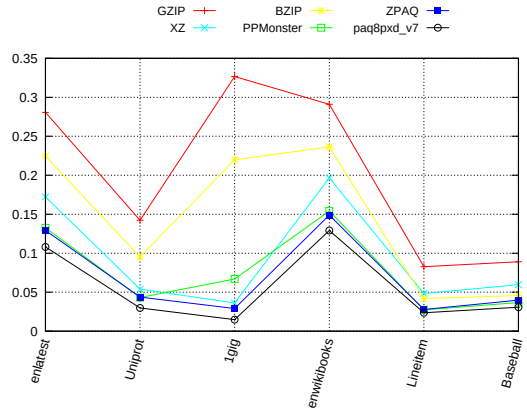


Figure 10: Compression of entire XML document with single instances of Vanilla Compressors.

Algorithm 4: Annotated Process.

```
1:  function ENCODE(AnnotatedTree A, File file)
2:  |  write(encode(A.schema), file.schema)
3:  |  DepthFirstIterator dfs
4:  |  // For each node, its annotation list.
5:  |  for (dfs = A.iterator(); dfs.hasNext();) do
6:  |  |  AnnotatedNode node = dfs.next()
7:  |  |  write(node.annotationList.length(), file.annot).
8:  |  |  write(node.annotationList, file.annot).
9:  |  end for
10: |  // For each node write its text container.
11: |  for (dfs = A.iterator(); dfs.hasNext();) do
12: |  |  AnnotatedNode node = dfs.next()
13: |  |  write(node.textContainer.length(), file.text).
14: |  |  write(node.textContainer, file.text).
15: |  end for
16: end function
```

[1] gzip -9 FILE
[2] xz -9 -e FILE
[3] bzip2 -9 FILE

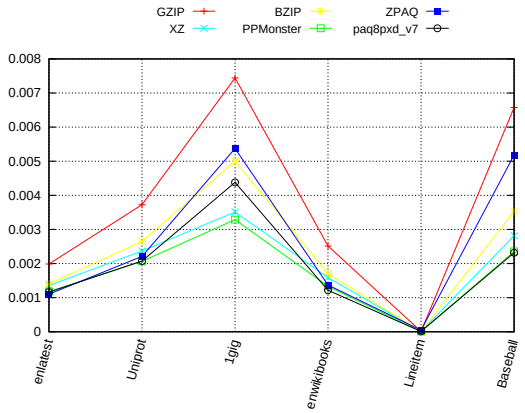


Figure 11: Compression of Annotated Tree (the number of bytes to represent the XML Syntax/Structure) over Markup Density.

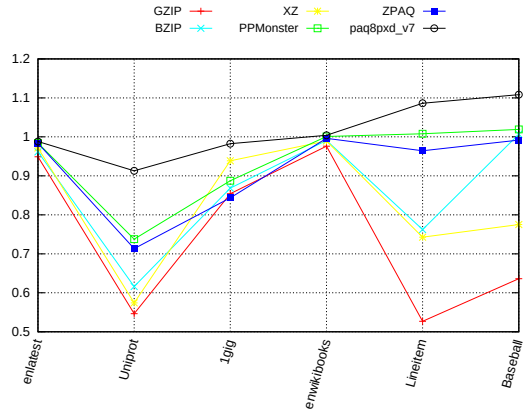


Figure 12: Compression of Annotated Transform (collection of Annotated Tree, Text Container, and Schema Tree) over Compression of XML Data.

tion by Partial Matching based PPMonster[4], and the Context-Mixing compressors ZPAQ (ZPAQ, 2013)[5] and PAQ8PXD_V7[6]; for the source code, or executables, to each compressor see (Mahoney, 2012). Results of tests showing applications of the annotated transform to each document and compressing only the structure ($\frac{\text{Annotations + Annotated Tree}}{\text{Markup Density}}$) are shown in Figure 11. It represents the syntax of the each XML document in a fraction a percent. The annotation list of each equivalence class has two components: (1) A single byte header, that signifies if any transform (e.g., a run length encoding) has been applied to the annotations; (2) A list of annotations, each encoded as a 32-bit integer (although this can be much improved by using variable sized bytes). In the worst case, the annotated representation only requires one tenth of a percent of the original markup amount (including tag names, and XML syntax data). In the best case, a very-regular (a complete tree) document, lineitem, only requires one one-thousandth of a percent of the original markup amount. In either situation, both of these situations offer a very faithful yet succinct representation of the XML data.

Finally, Figure 12 plots the compression ratio of the size of the annotated tree over the compression ratio of the XML document shown in Figure 10. The first noticeable feature of Figure 12 is the fact that paq8pxd and PPMonster compress the data much better as vanilla compressors than with the annotated transform for the smaller XML files. Since the files are so small, these compressors can often build a model of the entire document, allowing those compressors to compress each tag-name, and the XML markup, quite compactly. In addition, by merging all of the text-containers into one compressible document, the data among container boundaries will often harm the initial compression of the subsequent container (the internal models have to adapt). The next general trend shown in Figure 12 is that the more markup-dense XML documents receive the more optimal compression ratios, whereas the more content-dense XML documents, only receive slight improvements for the larger XML documents. With respect to enlatest and enwikibooks, the majority of text is free-formed English, e.g., each `<text>` tag contains a substantial amount of text data (all of the content you would see on a Wikipedia page). If some nodes text data were of significant size, only the compressors that incorporate a very large scope of the data would be able to exploit the tag-to-tag redundancy, otherwise, it would only be able to exploit the redundancy local to that tag, and the data at the boundary of two tags. From this, we can infer that the scope of compression is a necessary and sufficient factor in the performance of compression shown in Figure 10 (i.e., because a larger scope allows better compression of the XML syntax and the semantic/temporal relations among subtrees). Another factor may be attributed to the fact that the lower bound of lossless compression for these documents is "close" to the obtained compression ratios.

6 CONCLUSIONS AND FUTURE WORK

This paper showed that annotated trees form a faithful representation of the trees, and so the XSAQCT compression process is lossless. The formal approach and

[4] ppmonstr -m1700 -o64 FILE
[5] zpaq add FILE.zpaq FILE -method 69 -noattributes
[6] paq8pxd_v7 -8 FILE

specific algorithms have both been provided. Besides the formal and algorithmic approaches, experiments showed that the annotated tree compressibility, without using any backend compressors is high, on average approximately 0.4. Finally, a general analysis and results of testing of compression of entire XML document with single instances of vanilla compressors, compression of annotated tree over markup density, and compression of annotated transform over compression of XML data were provided, showing the usefulness of the annotated tree approach.

Simple queries, such as finding all children of a given node can be efficiently evaluated using the annotated trees. Our future work will extend queries to the subset of XPath expressions known as the core XPath as defined in (Gottlob et al., 2005), as well as more sophisticated navigational queries, e.g. asking for the j-th level-ancestor of u.

ACKNOWLEDGEMENTS

The work of the first and third authors are partially supported by the NSERC RGPIN grant and NSERC CSG-M (Canada Graduate Scholarship-Masters) grant respectively.

REFERENCES

Arion, A., Bonifati, A., Manolescu, I., and Pugliese, A. (2007). XQueC: a query-conscious compressed XML database. *ACM Transactions on Internet Technology*, 7(2).

Baseball.xml (2013). baseball.xml, retrieved October 2013 from http://rassyndrome.webs.com/cc/baseball.xml.

Benoit, D., Demaine, E., Munro, J., and Raman, V. (1999). Representing Trees of Higher Degree. In Dehne, F., Sack, J., Gupta, A., and Tamassia, R., editors, *Algorithms and Data Structures*, volume 1663 of *Lecture Notes in Computer Science*, pages 169–180. Springer Berlin Heidelberg.

Bille, P., Gortz, I., Weimann, O., and Landau, G. M. (2013). Tree Compression with Top Trees. In *In Proceedings of the 40th International Colloquium on Automata, Languages, and Programming*.

Burrows, M. and Wheeler, D. (1994). A block-sorting lossless data compression algorithm. *Technical Report, Digital Equipment Corporation*.

Busatto, G., Lohrey, M., and Maneth, S. (2005). Efficient Memory Representation of XML Documents. In Bierman, G. and Koch, C., editors, *Database Programming Languages*, volume 3774 of *Lecture Notes in Computer Science*, pages 199–216. Springer Berlin Heidelberg.

Busatto, G., Lohrey, M., and Maneth, S. (2008). Efficient memory representation of XML document trees. *Inf. Syst.*, 33(4-5):456–474.

bzip2 (2013). bzip2 compression, retrieved October 2013 from http://www.bzip.org/.

Chen, S. and Reif, J. (1996). Efficient Lossless Compression of Trees and Graphs. In *In IEEE Data Compression Conference (DCC)*.

Consortium, T. U. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. http://dx.doi.org/10.1093/nar/gks1068. Retrieved on June 20, 2013.

Corbin, T., Müldner, T., and Miziołek, J. (2013). Pre-order Compression Schemes for XML in the Real Time Environment. In *The Ninth International Conference on Web Information Systems and Technologies*, Aachen, Germany. WEBIST.

Corpus, W. (2013). Wratislavia XML corpus, retrieved October 2013 from http://www.ii.uni.wroc.pl/ inikep/research/wratislavia/.

Ferragina, P., Luccio, F., Manzini, G., and Muthukrishnan, S. (2009). Compressing and indexing labeled trees, with applications. *J. ACM*, 57(1):4:1–4:33.

Gottlob, G., Koch, C., and Pichler, R. (2005). Efficient algorithms for processing xpath queries. *ACM Trans. Database Syst.*, 30(2):444–491.

GZIP (2013). The gzip home page, retrieved October 2013 from http://www.gzip.org.

Jacobson, G. (1989). Space-efficient static trees and graphs. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, SFCS '89, pages 549–554, Washington, DC, USA. IEEE Computer Society.

Mahoney, M. (2012). Large Text Compression Benchmark, Retrieved October 2013 from http://mattmahoney.net/dc/zpaq.html.

Müldner, T., Corbin, T., Miziołek, J., and Fry, C. (2012). Design and Implementation of an Online XML Compressor for Large XML Files. *International Journal On Advances in Internet Technology*, 5(3):115–118.

Müldner, T., Fry, C., Miziołek, J., and Durno, S. (2009). XSAQCT: XML queryable compressor. In *Balisage: The Markup Conference 2009*, Montreal, Canada.

XML (2013). Extensible markup language (XML) 1.0 (Fifth edition), retrieved October 2013 from http://www.w3.org/tr/rec-xml/.

xmlgen (2013). The benchmark data generator, retrieved October 2013 from http://www.xml-benchmark.org/generator.html.

Ziv, J. and Lempel, A. (2006). A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theor.*, 23(3):337–343.

ZPAQ (2013). Zpaq, retrieved October 2013 from http://www.w3.org/tr/rec-xml/.

Filter-enabled Binary XML Dissemination in Embedded Networks

Sebastian Käbisch[1,2] and Richard Kuntschke[1]
[1]*Siemens AG, Corporate Technology, 81730 Munich, Germany*
[2]*University of Passau, Chair of Distributed Information Systems, 94032 Passau, Germany*
{*Sebastian.Kaebisch, Richard.Kuntschke*}*@siemens.com*

Keywords: Binary XML, EXI, Embedded Networks, XPath, Filtering

Abstract: Efficient data dissemination in distributed systems is a challenge that can be tackled by sharing common data and processing results among multiple queries. Doing so in an effective manner helps to save network bandwidth and computational resources. This is especially important in embedded networks where such resources are often extremely scarce. Disseminating resource-intensive XML data in embedded networks has been enabled by using binary XML technologies such as W3C's EXI format. In this paper, we show how filter-enabled binary XML dissemination in embedded networks helps to further reduce resource demands. Thus, through the suitable placement of pre- and post-filters on binary XML data, bandwidth on network connections and computational resources on nodes can be saved. Consequently, more data can be processed with a certain amount of available resources within an embedded network.

1 INTRODUCTION

Efficient data dissemination is an important task in distributed systems where data needs to be transferred from data sources to data sinks located at different places within the system. Sharing common data and processing results among multiple data sinks helps to save network bandwidth and computational resources. While such approaches reduce resource usage in any distributed system, their usage is crucial in embedded networks (fundamental in, e.g., building or industrial automation, automotive industry, and smart grid) due to the strictly limited resources such as from microcontrollers. Using binary XML techniques such as W3C's Efficient XML Interchange (EXI) format (Schneider and Kamiya, 2011) enables the dissemination of otherwise resource-intensive XML data in embedded networks. Additionally, using filter-enabled binary XML dissemination helps to further reduce resource demands.

A filter-enabled subscription mechanism in embedded networks reduces network traffic and unnecessary message processing at the client nodes. It optimizes data interaction between the service provider (data source) and the service requester/client (data sink) in terms of data novelty and supports an efficient execution of applications in embedded networks at runtime. Consequently, more data can be processed with a certain amount of available resources within an embedded network.

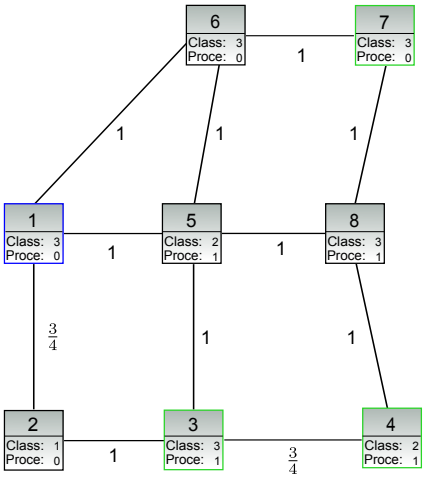


Figure 1: Example embedded network with a service provider (node 1) and three clients (node 3, 4, and 7)

An immediate evaluation at the node of service data origin prevents the dissemination of data that is outside of the scope of the corresponding service requesters. In the context of constrained embedded networks, however, a desired early evaluation at the node of service data origin sometimes cannot be realized. Two aspects support this observation: First of all, the available resources, especially when it comes to memory, are often not sufficient for installing an additional filter application. Secondly, vendors of em-

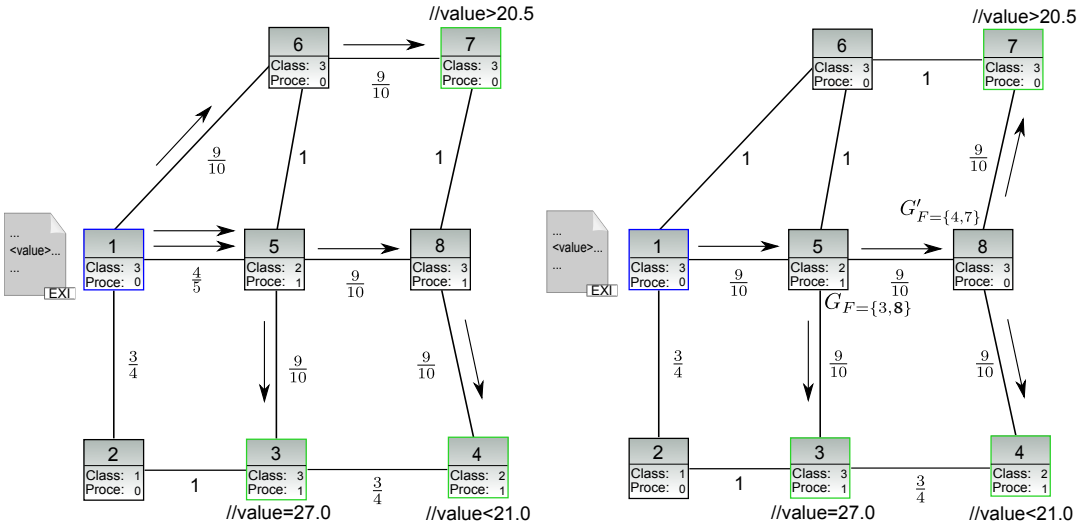


Figure 2: (a) Naive approach for data dissemination (b) Desired pre-/post-filter approach for data dissemination purposes.

bedded nodes do not offer the possibility of installing such filter applications.

In this paper, we introduce an approach for organizing filter-based service data dissemination in constrained embedded networks that takes into account resources such as device classes, device processing performance, and connection quality between nodes. The goal is to share relevant service data using binary XML whenever possible by using filters and subfilters, respectively. This reduces both network traffic and computational load on nodes. Basically, this approach leads to content-based routing at the application level based on binary XML content.

1.1 Problem Statement

Consider Figure 1 as an example embedded network that shows a simple distributed system with eight nodes and corresponding network connections. Nodes are categorized here into three classes (1 to 3) indicating their capabilities. The lower the class number, the more powerful the device. Class 1, e.g., corresponds to a consumer PC, class 2 corresponds to a hardware system with a resource complexity found in home network routers, and class 3 corresponds to constrained embedded hardware such as a microcontroller (e.g., ARM Cortex-M3[1] with 24MHz, 256kB RAM, and 16kB ROM). The boolean *proce* value indicates whether there are computational resources available at the corresponding node. A value of 0 indicates that there are no resources left for installing a potential application, while a value of 1 indicates available resources. Resources in this context may comprise memory as well as processing power. The numbers at the network connections indicate the available connection quality. Values closer to 1 indicate superior connection quality. In contrast, values closer to 0 indicate poor connection quality. Connection quality may be impacted by the technology of the underlying physical communication link, e.g., low-power wireless communication (e.g., IEEE 802.15.4 (IEEE, 2011)) vs. Ethernet, as well as by the current system state as indicated, e.g., by the current package loss ratio, delay, and available bandwidth of the communication link. Increasing network traffic on a communication link lowers the connection quality since, e.g., increasing traffic reduces the available bandwidth.

Now, let us assume a simple data dissemination scenario. In Figure 2(a), node 1 is the data source, sending XML-based data including a *value* information in EXI format, and nodes 3, 4, and 7 are the data sinks that are interested in the *value* information. The data sink at node 3 is only interested in data elements containing the value 27.0, node 4 is interested in data elements containing values less than 21.0, and node 7 is interested in data elements containing values greater than 20.5. Since node 1 does not have the opportunity to set up a filter mechanism (*proce* is equal to 0) to test client's relevance of a new data message, this node always sends an individual copy of the entire data message to each data sink and the nodes constituting the data sinks process the data accordingly (see Figure 2(a)). However, this leads to relatively high bandwidth usage on the affected network connections and requires considerable computational re-

[1] http://www.arm.com/products/processors/cortex-m/cortex-m3.php

sources at the data sink nodes, even if this message is not relevant anyway. To illustrate the impact of this naive dissemination approach in terms of connection quality, in Figure 2(a) in each involved dissemination link the connection quality value is decremented by the value $\frac{1}{10}$.

Now consider Figure 2(b), showing our pre-/post-filter approach for the same distributed network. All produced data is sent out by the data source only once. A pre-filter (G_F) at node 5 filters the data of relevance for nodes 3 and 8. A message for node 3 is only forwarded when the *value* in the data message is equal to 27.0. Node 8 receives only a message from node 5 when a *value* information is present in the message. A post-filter (G'_F) at node 8 is applied to evaluate the relevance for nodes 4 and 7. Consequently, node 8 forwards only the messages to node 4 when the *value* is smaller than 21.0. Node 7 only receives a message when the *value* is greater than 20.5.

Thus, compared to the naive approach, the requested data is transmitted to each data sink using less network bandwidth and less processing power overall in the system, since processing results can be shared among multiple data sinks. Also, we are able to distribute the computational load across capable nodes within the distributed system. We accept this approach even if the resources of a small number of node devices may be claimed by the filter placement. This is seen by the processablity of node 8 that is set to 0.

1.2 Paper's Contributions and Outline

This paper presents the following contributions:

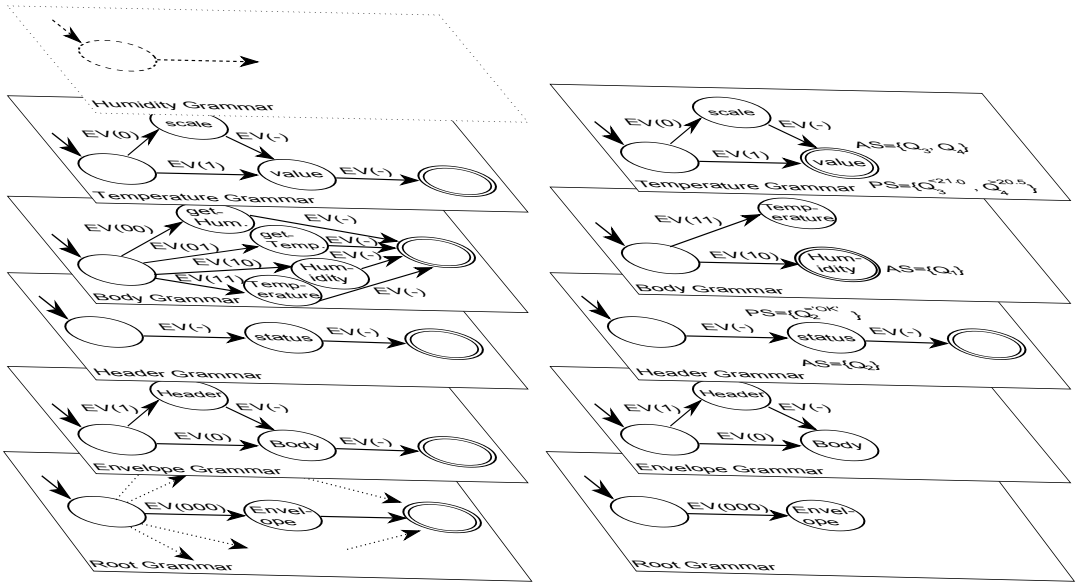
- We shortly introduce a previously developed filter mechanism based on W3C's EXI format (Section 2). This mechanism constitutes the basis for our further work on filter-enabled binary XML dissemination in embedded networks.
- In Section 3, we present our new approach for efficient filter-enabled binary XML data dissemination in embedded networks. The approach helps to reduce bandwidth usage on network connections and computational load on nodes, thus allowing larger amounts of data to be transferred and processed with a certain amount of available resources.
- Finally, we present evaluation results showing the effectiveness of our approach (Section 4).

2 EXI FILTER MECHANISM

In this section we give a short introduction about the EXI format and the basic idea, how an EXI grammar can be transformed to a filter grammar for evaluating XPath expressions.

2.1 The W3C EXI Format

W3C, the inventor and standardizer of XML, faced the drawbacks of plain-text XML and created a working group called XML Binary Characterization (XBC) (Goldman and Lenkov, 2005) to analyze the condition of a binary XML format that should also harmonize with the standardized plain-text XML format as well as with the XML Infoset. The outcome was the start of the W3C Efficient XML Interchange (EXI) format, which gained recommendation status at the beginning of 2011 (Schneider and Kamiya, 2011). Mainly, EXI is a grammar driven approach that is applied to bring XML-based data into a binary form and vice versa. Such a grammar is constructed based on a given XML Schema where each defined complex type is represented as a deterministic finite automaton (DFA). Figure 3(a) shows an excerpt of a sample EXI grammar (set of automaton) G that can be used for encoding and decoding. This grammar reflects an XML Schema including the SOAP framework (Gudgin et al., 2003) with a status information within the Header part and request/response patterns of temperature and humidity information within the Body part. Please note that the *Root* grammar is a predefined grammar that occurs in each EXI grammar representation of arbitrary XML Schemas. It contains all entry points of all root elements in a given schema. Here, we only highlight in our context the relevant root Envelope of the SOAP message framework. In general, each DFA contains one start state and one end state, which reflect the beginning and the end, respectively, of a complex type declaration. Transitions to the next state represent the sequential order of element and/or attribute declarations within a complex type. Optional definitions (e.g., *choice*, $minOccurs = 0$, etc.) are reflected by multiple transitions and assigned an event code (EV). E.g., the SOAP Envelope message framework embeds an optional Header and a mandatory Body element (Gudgin et al., 2003). The equivalent EXI grammar representation, as can be seen in Figure 3(a) (Envelope Grammar), provides two transitions from the start state: one to the Header state and one to the Body state. For signalization, a one bit event code is used and assigned to the transition (EV(1) for the Header; EV(0) for the Body). Generally, the number of bits used for m transitions is deter-



(a) (b)

Figure 3: (a) EXI grammar G (excerpt) for encoding and decoding purposes (b) Filter Grammar G_F with $G_F \subseteq G$ based on the queries Q_1, Q_2, Q_3, and Q_4 for evaluation purposes.

mined by $\lceil log_2 m \rceil$. EV(-) on transitions indicates, no event code is required.
An example XML message snipped such as

<Envelope><Header><status>OK< /status>...

would be transformed to a

000 1 'OK' ...

EXI (bit-)stream based on the EXI grammar shown in Figure 3(a). This already shows, how compact EXI can be. In addition, EXI is a type aware encoder that provides efficient coding mechanisms for the most common data types (int, float, enumerations, etc.). There are use cases in which the EXI representation is said to be over 100 times smaller than XML (Bournez, 2009). Based on the high compression ratio and the opportunity to obtain the data content directly from the EXI stream, XML-based messaging is also feasible in the embedded domain, even if constrained devices are used (Käbisch et al., 2011).

2.2 EXI Filtering

EXI grammars build the bases for writing and reading binary XML data. Previous work (Käbisch et al., 2012) shows the functionality to create an efficient filter mechanism for binary XML data based on a number of service requesters by providing XPath expressions that address the desired service data occurrences and/or data value conditions. Two approaches which are feasible to constrained devices such as microcontrollers were presented: *BasicEXIFiltering* and *OptimizedEXIFiltering*. The *BasicEXIFiltering* operates on top of an EXI grammar and evaluates normalized XPath queries by means of binary XML. *OptimizedEXIFiltering* presents a more sophisticated approach; it maps all XPath expressions within an EXI grammar and removes all states and transitions which are not required for message evaluation. The outcome is a filter grammar denoted as G_F. Figure 3(b) shows such a constructed filter grammar G_F based on the queries

$$Q_1 = //Humidity$$
$$Q_2 = //status[text() =' OK']$$
$$Q_3 = //Temperature/value[text() < 21.0]$$
$$Q_4 = //Temperature/value[text() > 20.5]$$

applied on the data model represented by the EXI grammar G in Figure 3(a).
Applying the previously created EXI stream

000 1 'OK' ...

to the filter grammar G_F we would, at least, successfully find a match for the XPath expression Q_2. This is due to the remaining states and transitions in G_F that lead via the Envelope state (reading the bits 000 from the stream) and the Header state (reading the bit 1 from the stream) to the Header status state which is dedicated as a predicate state (PS) and as an accepting state (AS). A predicate state results to a predicate evaluation. Here, we have to evaluate whether the message contains the status value 'OK' which is also the case in our example. An accepting state rep-

resents a query match of a particular query (here Q_2). This is only true, if all aforegoing relevant predicate states resulted in a positive predicate evaluation.

The next section explains how such filter grammars can be used to realize an efficient binary XML data dissemination mechanism in constrained embedded networks.

3 DISSEMINATION ALGORITHM

This section presents our filter-enabled binary XML dissemination algorithm. First of all, we introduce our cost model and formalize the optimization problem (Section 3.1). Next, the dedicated algorithm is presented (Section 3.2). Finally, an example is considered for better clarification of how our algorithm works (Section 3.3).

3.1 Cost Model

Before we formalize our cost model we will formally define an embedded network in our context as described in Section 1.1 by $N_{emb} = (V, E, c, w, p)$. V describes the set of vertexes/nodes. Set E describes the set of edges/connections between two nodes. Function c with $c : V \to \mathbb{N}_{>0}$ associates the device class of a device node. The weight function w with $w : E \to \mathbb{R}_{[0,1]}$ describes the connection quality between two device nodes and p with $p : V \to \{0,1\}$ the processing capability of a node.

A newly installed application, including a service provider node (v_s) and service subscriber clients ($C = \{v_{c_1}, ..., v_{c_n}\}$), in an embedded network N_{emb} would typically lead to additional network traffic and processing costs. To keep this overhead as small as possible, we filter the data of relevance and share this data as long as possible on a determined dissemination path that avoids constrained device class nodes and uses connections with relatively good quality. Consequently, we have two metrics which have to be considered: device class with the processability (c and p) and connection quality (w). Putting this together, we define our cost function f for a given $N^T_{emb} = (V^T, E^T, c, w, p)$ with $N^T_{emb} \subseteq N_{emb}$ that only contains the nodes and transitions (spans a tree) from v_s to all nodes in C that is used for the data dissemination:

$$f(N^T_{emb}) := \alpha \cdot \sum_{v_i \in V^T} (c(v_i) + 1 - p(v_i)) + (1-\alpha) \cdot \sum_{(v_i,v_j) \in E^T} \frac{1}{w(v_i, v_j)} \quad . \tag{1}$$

The first summand formalizes the device class with the processing capability of a device node. A *hop-noise* value 1 is added to enable an influential decision when we also have only one class occurrences ($c = 1$) and each has processing capability ($p = 1$) in the network. The second summand represents the reciprocal connection quality. The $\alpha \in [0,1]$ is a weight factor that enables us to set up a more dominant part in the cost function: the device class ($\alpha > 0.5$) or the connection quality ($\alpha < 0.5$).

Using f we are able to formalize our optimization problem to find a subgraph N^T_{emb} of N_{emb} for a filter-enabled service data dissemination:

$$Minimize \quad f_T \tag{2}$$

subject to

$$\sum_{v_i \in V^T} p^T(v_i) \geq 1 \quad .$$

The inequality constraint specifies the occurrence of at least one node processing capability within N^T_{emb} that can be used to set up a pre-filter.

Unfortunately, for any constellation in N_{emb} we are not able to find an optimized solution in polynomial time. In the next section we are going to present an heuristic approach based on greedy algorithms that approximates an optimized N^T_{emb} for a filter-enabled service data dissemination.

3.2 Algorithm

We are now going to describe our filter-enabled service data dissemination algorithm, the *FilterEnabledDissemination* algorithm (see Algorithm 1), for installing a new application with a service provider and a number of service requesters, which takes into account the current resources of the embedded network.

As input, the algorithm takes an embedded network N_{emb}, a dedicated service provider node v_s, a set of service requesters (the clients) C and their corresponding queries Q, and the underlying data model of the service provider described in an XML schema *XSD*. Its outcome is a subnetwork N^T_{emb} of N_{emb} that represents the dissemination tree/path from v_s to all clients in C and a set F that consists of the selected nodes with pre- and post-filter properties. Essentially, the processing steps of the algorithm can be divided into three parts:

1. After determining the filter grammar G_F (see Section 2.2) by the *FilterGrammar* in line 1, a suitable pre-filter node is searched. Doing this, the *ClosestPreFilterNode* algorithm (line 2) is

Algorithm 1: *FilterEnabledDissemination*.

Require: $N_{emb} = (V, E, c, w, p)$, a service provider v_s, set of service requesters $C = \{v_{c_1}, ..., v_{c_n}\}$, set of queries Q related to client's conditions, data model represented as XML schema *XSD*.

Ensure: Tree network N^T_{emb} with a set F of dedicated selected nodes with its filter grammars (pre- and post-filters).

1: $G_F \leftarrow FilterGrammar(G, Q)$;
2: $v_{pre} \leftarrow ClosestPreFilterNode(N_{emb}, v_s, C, G_F)$;
3: $N^T_{emb} \leftarrow DisseminationTree(N_{emb}, v_{pre}, C)$;
4: $F \leftarrow PostFilterPlacement(N^T_{emb}, Q, v_{pre}, G_F)$;
5: $extendTreeByPreRoute(V^T, E^T, V, E, v_{pre})$;
6: **return** $\{N^T_{emb}, F\}$

called. This algorithm determines one processing node v_{pre} that is able to run the filter grammar G_F as well as that results in overall positive data dissemination. More precisely, we are not only considering the quality of the path to a processable node v_{pre} in terms of connection and device class, but also the quality from v_{pre} to all service subscribers.

2. Starting with the determined pre-filter node v_{pre} we discover an optimized dissemination tree N^T_{emb} from this v_{pre}. The *DisseminationTree* algorithm (line 3) will be called to gather such a tree. Thereby, relevant service data shall be delivered from v_{pre} to the service subscribers in a resource-optimized manner. More precisely, the data shall be routed via high quality connections, avoid very constrained embedded devices, and be shared for as long as possible if there are multi-client destinations. The latter can be fulfilled if one or more post-filters can be placed that retain the information of the final client destination nodes or the next post-filter nodes. Consequently, the *DisseminationTree* algorithm finds a dissemination tree from v_{pre} to all clients in C that takes into account the device class and connection quality metrics as well as the current processing capability of the potential post-filter placement.
3. Based on N^T_{emb}, suitable nodes are selected for the post-filter functionality to share service data as long as possible. The *PostFilterPlacement* algorithm (line 5) realizes this and provides the routing information for all filter grammars.

Before the *FilterEnabledDissemination* algorithm terminates, we extend N^T_{emb} by the involved nodes and connection that leads from v_s to v_{pre} (line 5).

For better clarification and to get an idea how our algorithm works we will consider an example in the next subsection applied on the network shown in Figure 1.

3.3 Example

We will now consider the embedded network $N_{emb} = (V, E, c, w, p)$ which is shown in Figure 1. Node 1 is a service provider and nodes 3, 4, and 7 are the service requesters with the following 4 query conditions as presented in Section 2.2:

Node 3: Q_1 and Q_2
Node 4: Q_3
Node 7: Q_4

Figure 3(b) already shows the filter grammar G_F based on this query set after applying the *FilterGrammar* procedure. Since the service requester node does not provide us with the opportunity to set up a filter mechanism for clients' subscription requests ($p(1) = 0$), we have to find an alternative node for placing a pre-filter. In order to do so, we have to identify any nodes with processing capabilities within the network that have enough resources to run G_F. The procedure *ClosestPreFilterNode* in Algorithm 1 will identify these nodes (3, 4, 5, 8) check their resource capabilities. If this results in more than one node, the node which is closest to the service provider and yields the best cost function value is selected based on test paths to clients in C using the Dijkstra algorithm (Dijkstra, 1959) with our metrics. The outcome would be node 5 that is selected as pre-filter node (v_{pre}) running G_F.
The next step involves determining a data dissemination tree that spans v_{pre} and client nodes 3, 4, and 7. In order to do so, we will call the *DisseminationTree* procedure. This algorithm is based on the concept of the Kou-Markowsky-Berman (KMB) algorithm (Kou et al., 1981) which is a well-known heuristic for the Steiner Tree problem. Based on our metric and cost model, respectively, Figure 4 shows the outcome of the *DisseminationTree* procedure.
The last major processing step in our dissemination algorithm involves determining suitable post-filter nodes to enable a high ratio of shared service data from service provider to service requesters. Starting with (root) node 5, the *PostFilterPlacement* procedure will first select all nodes that contain multi successor branches. Nodes 5 and 8 are candidates. Since node 5 already is a dedicated pre-filter node, we will not consider it further and instead check node 8 directly for processability of a post-filter grammar. The post-filter grammar is constructed based on the queries that can be reached from node 8. This is

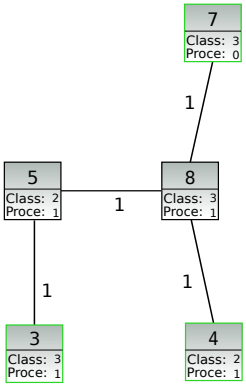


Figure 4: Determined tree based on the *DisseminationTree* procedure from v_{pre} (node 5) to all clients in C (nodes 3, 4, and 7).

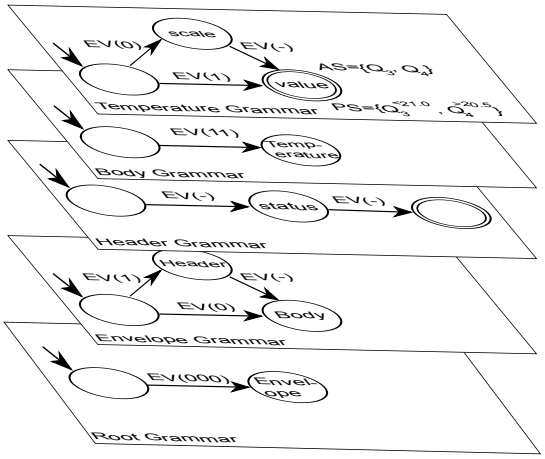


Figure 5: Post-filter grammar.

true for the queries Q_3 and Q_4. Figure 5 shows the post-filter as based on these queries that is constructed based on the mechanism presented in Section 2.2. Since node 8 is a node with processing capabilities we can successfully install the post-filter on this node.
Before the *PostFilterPlacement* algorithm is terminated, we are going to update the network's defined filters in terms of routing information. Node 5 is set up with the pre-filter G_F that is shown in Figure 3(b) and will receive all messages from node 1. In addition, node 5 with G_F contains the associated information Q_1 and Q_2 related to node 3, Q_3 related to node 4, and query Q_4 related to node 7. Based on the post-filter to be placed on node 8, service data that matches queries Q_3 and Q_4 shall be forwarded to node 8, which then will send the data only once. Thus, G_F is updated with this information. In summary, we obtain the following routing information:

- G_F (at node 5): forwards service data to node 3 when queries Q_1 and/or Q_2 match; forwards service data to node 8 when queries Q_3 and/or Q_4 match.
- G'_F (at node 8): forwards service data to node 4 when Q_3 matches; forwards service data to node 7 when Q_4 matches

This determined data dissemination tree and the filter placement is also reflected in Figure 2(b).

4 EVALUATION

So as to organize service data dissemination of each new applied application and to estimate its influence in terms of traffic and device capacity usage of real embedded networks we wrote an embedded network simulator. The simulator provides us with the opportunity to load particular network topologies and characteristics as well as service provider and the service subscribers with their queries. Another alternative is to setup randomized embedded networks by providing different kinds of generation parameters: number of nodes, number of different kinds of device classes, and the ratio of device classes and connection quality. Based on such a network, we are able to set up new applications by selecting particular nodes, which operate a service with the provided service description, and the client nodes that subscribe service data with the predefined conditions on the service data. We can then run our dissemination algorithm for each newly installed application.
In order to evaluate the effectiveness of the approach presented in this paper, we randomly generated an embedded network that has a complexity of 50 nodes with three device classes. This network setup initially features a balanced ratio of processable and non-processable nodes. Its class ratio consists of 5 times device classes 1, 10 times device classes 2, and 35 times device classes 3. Initially, we uniformly distributed the connection quality weighting values with numbers between 0.8 and 1. We sequentially installed five different kinds of applications. In general, an application is based on a service provider and different kinds of service requesters (the clients). The distance (in terms of hop count) and client distribution to the service provider node is increased with each new installed application. We start with the first application, which has two clients; subsequently, the second has 3 clients, the third has 4 clients, there are 5 clients in the fourth application, and finally the fifth application has 6 different service requesters. For each installed application we evaluated the service data dissemination for two variants: *Filter-enabled dissemination* (abbreviated with *FD*) represents our filter-enabled dissemination approach and the separate and *direct dissemination* (abbreviated with *DD*)

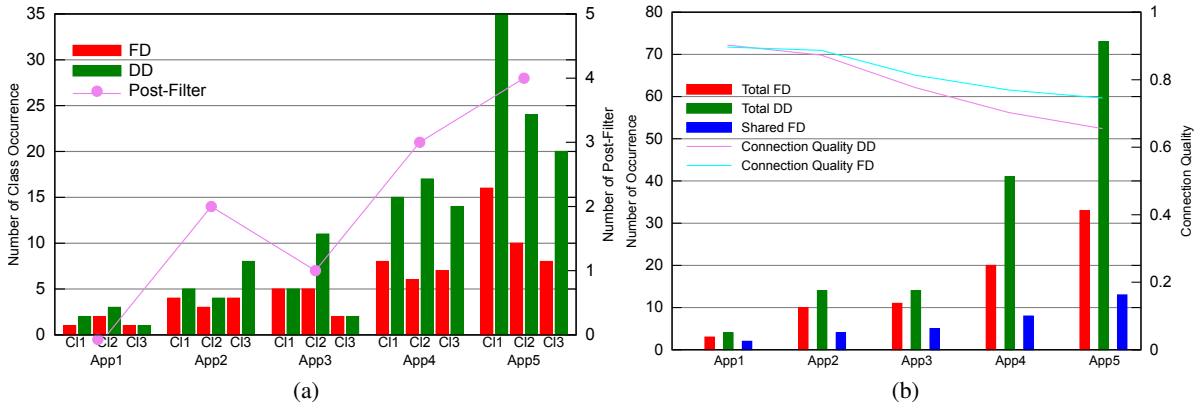


Figure 6: Embedded network with $|V| = 50$: (a) Count of used classes (Cl1, Cl2, Cl3) in the dissemination path and number of used post-filters for each application (App1, ..., App5). (b) Number of links of the dissemination path and average value of connection quality.

reflects the direct, non-filtered service data delivery (comparable with Figure 2(a)).

Figure 6 shows the evaluation results. Figure 6(a) depicts the result for each application in terms of device class occurrences (Cl1=Class 1 nodes, Cl2=Class 2 nodes, and Cl3=Class 3 nodes) in the dissemination path of our approach (FD Optimized) as compared to the simple approach, wherein each service data is delivered separately (DD Simple). In other words, we count the occurrence of the device classes in the determined dissemination path (tree) that reflects the worst case scenario when a service message is relevant for all service requesters in the network. As can be seen for all cases, our approach, as presented in this paper, results in a lower usage of class occurrences as compared to the simple service data distribution variant. This becomes especially apparent the more complex the application is. Furthermore, the occurrences also show that our determined dissemination paths always consist of the desirable, relatively small number of constrained nodes (class 3). For instance, in a worst case distribution scenario for application five, our dissemination approach uses the device class 1 sixteen times, class 2 ten times, and the most constrained device class 3 eight times. In total, 34 nodes are involved in the dissemination process. In contrast, a simple dissemination would lead to a device class ratio of class 1 thirty-five times, class 2 twenty-four times, and class 3 twenty times. In total, this involves 79 nodes. Consequently, our approach results in a better resource usage of the nodes in the embedded network since fewer total nodes are involved in the dissemination tree; the number of constrained nodes (class 3) is kept as small as possible.

Figure 6(b) shows the evaluation result in terms of the number of connection links used and average connection quality. As can be seen, the number of connections used in a dissemination process is smaller for our approach as compared to the simple variant. The figure also shows the ratio of the shared connections of the optimized variant in each application. We determined the number based on whether each connection between two nodes can reach a pre-filter or a post-filter node. If so, the number of shared connections is incremented. The presented numbers show the effectiveness of our approach, since for each application we determine a dissemination tree that consists of a high ratio of shared connections. Figure 6(b) also shows the average connection quality for each application and its dissemination based on both our approach and the simple variant. As can be observed, the simple dissemination variant loses the average connection quality faster than our approach. This is explained by the fact that the simple variant involves a lot more connection links and potentially causes more network traffic. The more applications that are installed in the network, the greater the impact on connection quality will be.

5 RELATED WORK

Finding a suitable pre-filter node outside of the service data origin node and the position of post-filters in a dissemination tree opens the opportunity to share relevant service data with a number of service subscribers. This leads to a reduction of resources used within embedded networks in terms of network traffic as well as processing overhead. Similar topics are addressed by and can be found in Data Stream Management Systems (DSMSs). DSMSs complement the traditional Database Management Systems (DBMSs). Typically, a DBMS handles persistent and random ac-

cessible data and executes volatile queries. Meanwhile, in DSMSs persistent (or long-running) queries are executed over volatile and sequential data. Examples of DSMSs include *Aurora* (Abadi et al., 2003), *Borealis* (Abadi et al., 2005), *TelegraphCQ* (Chandrasekaran et al., 2003), and *StreamGlobe* (Kuntschke et al., 2005). The main focus of such systems is on the efficient processing of potentially infinite data streams against a set of continuous queries. In contrast to publish/subscribe systems such as XFilter (Altinel and Franklin, 2000) or YFilter (Diao and Franklin, 2003), continuous queries in DSMSs can be far more complex than simple filter subscriptions. Some research developed new query languages such as WindowedXQuery (WXQuery) (Kuntschke, 2008) to extend query operations. In the domain of constrained embedded networks, however, we presume the presence of relatively simple data models and have found that XPath expressions are sufficient to address data interests and simple constraints by predicates. Other important work in distributed DSMSs such as StreamGlobe and Borealis revolves around network-aware stream processing and operator placement. These are issues also relevant to constrained embedded networks and, similarly, we took them into account for our approach by positioning the pre-filter and, if possible, the post-filter mechanism at the embedded nodes.

Most DSMSs, such as TelegraphCQ for example, are based on relational data. StreamGlobe, however, focuses on plain-text XML data streams as well as on XML-based query languages such as XQuery (Boag et al., 2007) or the above mentioned WXQuery. Consequently, nodes used for distributed data stream processing in systems such as StreamGlobe and Borealis generally need to be far more powerful than the microcontrollers for constrained embedded devices that we aim for in this paper. Our approach for constructing high performance filter mechanisms based on binary XML techniques enables us to bring DSMS topics to the domain of constrained embedded networks.

In our approach, filter nodes such as pre-filter or post-filter nodes decide how to best forward service messages if there are one or more matches. The destinations may include service requester nodes and/or other post-filter nodes. In the literature, this is called content-based routing or application-level routing since routing depends on the contents of data within a message. In that context, we can refer to works such as the *combined broadcast and content-based* (*CBCB*) routing scheme (Carzaniga and Wolf, 2002), the *application layer multicast algorithm* (*ALMA*) (Ge et al., 2006), the usage of *XML Router* (Snoeren et al., 2001), and *view selection for stream processing* based on XML data (Gupta et al., 2003). Below, we will concentrate on the latter since they also involve XML-based data content.

The XML Router approach (Snoeren et al., 2001) creates an overlay network that is implemented by multiple XML routers. An XML router is a node that receives XML packets and forwards a subset of these XML packets. The XML packets are forwarded to other routers or the final client node destinations. Thereby, the output links represent the XPath queries that describe the portion of the router's XML stream that should be sent to the host on that connection link. XML routers are comparable to our pre- and post-filter concept. However, additional strategies, such as reassembling a data packet stream from diverse senders provided by the diversity control protocol (DCP) or the usage of plain-text XML and XPath interpreters are not feasible in a resource constrained embedded environment.

The view selection for stream processing method is an interesting approach followed in (Gupta et al., 2003) and (Gupta et al., 2002). The main concept includes selecting a set of XPath expressions which are called *views*. The service data producers evaluate the views and add the result to the data package in the form of a header. The advantage is that servers which keep a local set of queries can evaluate their workload by inspecting only the values in the header and do not need to parse the XML document. This leads to a speed-up of routing decisions. However, this is only true for cases in which the evaluation in the header is positive. Otherwise, the complete (plain-text) XML document needs to be parsed and the query needs to be evaluated in a conventional way. Again, this is an obstacle in the constrained embedded environment. In addition, one of our goals is to achieve seamless protocol usage and to work with standardized message representations to support interoperability in a heterogeneous network environment. Adding a header to a message would break this principle and necessitate an adjustment of communication protocols.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented an approach to realize efficient filter-enabled service data dissemination in constrained embedded networks based on XPath expressions given by different subscribers/clients. Finding a suitable pre-filter node in an embedded network leads to an early evaluation of relevant service messages. By using post-filters in a determined dissemination tree, we are able to avoid redundant

transmissions and share the service data, especially if there is a multi-query match of different kinds of service requesters. The effectiveness in terms of device class occurrences, connection quality, and number of shared connections was demonstrated in a simulated environment based on our embedded network simulator.

Topics for future work include the dynamic update of client queries and their impact on the dissemination path as well as on the placed pre- and post filter grammar.

REFERENCES

Abadi, D. J., Ahmad, Y., Balazinska, M., Çetintemel, U., Cherniack, M., Hwang, J.-H., Lindner, W., Maskey, A., Rasin, A., Ryvkina, E., Tatbul, N., Xing, Y., and Zdonik, S. B. (2005). The design of the borealis stream processing engine. In *CIDR*, pages 277–289.

Abadi, D. J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., and Zdonik, S. B. (2003). Aurora: a new model and architecture for data stream management. *VLDB J.*, 12(2):120–139.

Altinel, M. and Franklin, M. J. (2000). Efficient filtering of xml documents for selective dissemination of information. In Abbadi, A. E., Brodie, M. L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., and Whang, K.-Y., editors, *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 53–64. Morgan Kaufmann.

Boag, S., Chamberlin, D. D., Fernández, M. F., Florescu, D., Robie, J., and Siméon, J. (2007). XQuery 1.0: An XML query language. World Wide Web Consortium, Recommendation REC-xquery-20070123.

Bournez, C. (2009). Efficient XML interchange evaluation. W3C working draft, W3C. http://www.w3.org/TR/2009/WD-exi-evaluation-20090407.

Carzaniga, A. and Wolf, A. L. (2002). Content-based networking: A new communication infrastructure. *Lecture Notes in Computer Science*, 2538:59–??

Chandrasekaran, S., Cooper, O., Deshpande, A., Franklin, M. J., Hellerstein, J. M., Hong, W., Krishnamurthy, S., Madden, S., Raman, V., Reiss, F., and Shah, M. A. (2003). Telegraphcq: Continuous dataflow processing for an uncertain world. In *CIDR*.

Diao, Y. and Franklin, M. J. (2003). High-performance xml filtering: An overview of yfilter. *IEEE Data Eng. Bull.*, 26(1):41–48.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Ge, M., Krishnamurthy, S. V., and Faloutsos, M. (2006). Application versus network layer multicasting in ad hoc networks: the ALMA routing protocol. *Ad Hoc Networks*, 4(2):283–300.

Goldman, O. and Lenkov, D. (2005). XML binary characterization. World Wide Web Consortium, Note NOTE-xbc-characterization-20050331. http://www.w3.org/TR/2005/NOTE-xbc-characterization-20050331.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., and Frystyk Nielsen, H. (2003). SOAP version 1.2 part 1: Messaging framework. World Wide Web Consortium, Recommendation REC-soap12-part1-20030624.

Gupta, A. K., Halevy, A. Y., and Suciu, D. (2002). View selection for stream processing. In *WebDB*, pages 83–88.

Gupta, K., Suciu, and Halevy (2003). The view selection problem for XML content based routing. In *PODS: 22th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*.

IEEE (2011). IEEE Standard for Local and metropolitan area networks–Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs). http://standards.ieee.org/about/get/802/802.15.html.

Käbisch, S., Kuntschke, R., Heuer, J., and Kosch, H. (2012). Efficient Filtering of Binary XML in Resource Restricted Embedded Networks. In *Proc. of the 8th International Conference on Web Information Systems and Technologies (WEBIST 2012)*, pages 174–182.

Käbisch, S., Peintner, D., Heuer, J., and Kosch, H. (2011). Optimized XML-based Web Service Generation for Service Communication in Restricted Embedded Environments. 16th IEEE International Conference on Emerging Technologies and Factory Automation.

Kou, Markowsky, and Berman (1981). A fast algorithm for steiner trees. *ACTAINF: Acta Informatica*, 15.

Kuntschke, R., Stegmaier, B., Kemper, A., and Reiser, A. (2005). Streamglobe: Processing and sharing data streams in grid-based p2p infrastructures. In Böhm, K., Jensen, C. S., Haas, L. M., Kersten, M. L., Larson, P.-Å., and Ooi, B. C., editors, *VLDB*, pages 1259–1262. ACM.

Kuntschke, R. B. (2008). *Network-aware optimization in distributed data stream management systems*. PhD thesis, Technical University Munich. urn:nbn:de:bvb:91-diss-20070806-625762-1-3; http://d-nb.info/989053113.

Schneider, J. and Kamiya, T. (2011). Efficient XML Interchange (EXI) Format 1.0. http://www.w3.org/TR/exi. W3C Recommendation 10 March 2011.

Snoeren, A. C., Conley, K., and Gifford, D. K. (2001). Mesh based content routing using XML. In *SOSP*, pages 160–173.

QoS-Aware Parameterized Semantic Matchmaking Framework for Web Service Composition

Salem Chakhar, Alessio Ishizaka and Ashraf Labib
Portsmouth Business School, University of Portsmouth, Portsmouth, U.K.
{*Salem.Chakhar, Alessio.Ishizaka, Ashraf.Labib*}*@port.ac.uk*

Keywords: Web service, Service Composition, Matchmaking, Quality of Service.

Abstract: The paper presents a parameterized and highly customizable semantic matchmaking framework. The matchmaking approach on which this framework is based distinguishes three types of matching: functional attribute-level matching, functional service-level matching, and non-functional matching. The functional matching permits to eliminate web services that fail to meet the user functional requirements. The non-functional matching permits to categorize web services instances into different ordered QoS classes. A series of algorithms are advertised for the different types of matching. These algorithms are designed to support a customizable matching process that permits the user to control the matched attributes, the order in which attributes are compared, as well as the way the sufficiency is computed for all matching types.

1 INTRODUCTION

Individual web services are conceptually limited to relatively simple functionalities modeled through a collection of simple operations. However, for certain types of applications, it is necessary to combine a set of individual web services to obtain *composite* web services (Chakhar et al., 2011). A crucial issue within web service composition is related to the selection of the most appropriate one among the different candidate web services.

In this paper we propose a semantic matchmaking framework for web service composition. We distinguish three types of matching: functional attribute-level matching, functional service-level matching, and non-functional matching. The functional attribute-level matching implies capability and property attributes and associates with each attribute a *similarity measure* that should be satisfied by the corresponding attribute in the advertised service. The functional service-level matching supports attribute-level matching and adds a service similarity measure that should be satisfied by the advertised service as a whole. The non-functional matching implies Quality of service (QoS) attributes.

The work presented in this paper is included in a layered system for web service composition that we proposed in (Chakhar, 2012). The matching operation intervenes in different levels of this system, as explained in Section 2.4.

The paper is structured as follows. Section 2 sets the background. Sections 3, 4 and 5 detail the framework. Section 6 provides illustrative examples. Section 7 presents experimental results. Section 8 discusses related work. Section 9 concludes the paper.

2 BACKGROUND

2.1 Example Scenario

This example scenario is freely inspired from a case use scenario described in the WSC Web Services Architecture Usage Scenarios (Hao et al., 2004). A company (travel agent) offers to people the ability to book complete vacation packages: plane/train/bus tickets, hotels, car rental, excursions, etc. Service providers (airlines, bus companies, hotel chains, etc.) propose web services to query their offerings and perform reservations. The user provides a destination and some dates to the travel agent service. The travel agent service inquires service providers about deals and presents them to the user.

2.2 Basic Definitions

We introduce some basic definitions of a service and other service-specific concepts. Several definitions are due to (Doshi et al., 2004).

Definition 1 (**Service**). *A service S is defined as a collection of attributes that describe the service. Let S.A denotes the set of attributes of service S and $S.A_i$ denotes each member of this set. Let S.N denotes the cardinality of this set.*

Example 1. *The travel agent company provides a web service,* ***bookVacation****, that is defined by the following attributes:* service category, input, output, preconditions, postconditions, response time, availability, cost, security, *and* geographical location.

Definition 2 (**Service Capability**). *The capability of a service S.C is a subset of service attributes ($S.C \subseteq S.A$), and includes only functional ones that directly relate to its working.*

Example 2. *The capability of* ***bookVacation*** *is:* $S.C = \{$input, output, preconditions, postconditions$\}$.

Definition 3 (**Service Quality**). *The quality of a service S.Q, is a subset of service attributes ($S.Q \subseteq S.A$), and includes all attributes that relate to its QoS.*

Example 3. *The Quality of* ***bookVacation*** *is:* $S.Q = \{$response time, availability, cost, security$\}$.

Definition 4 (**Service Property**). *The property of a service, S.P, is a subset of service attributes ($S.P \subseteq S.A$), and includes all attributes other than those included in service capability or service quality.*

Example 4. *The property of* ***bookVacation*** *is:* $S.P = \{$service category, geographical location$\}$.

2.3 Composition Approach

The key elements of the composition approach that we proposed in (Chakhar, 2012) are: the composition graph, potential executable plans and executable plan. The *composition graph* is an abstract representation of functional requirements provided by the user. It models the *invocation* relationships between the individual web services contained in the composite web service. The set of *potential executable plans* is the composite *service instances* which are obtained by replacing each service type in the composition graph by its instances using a set of transformation rules. The objective of the transformation operation is to include the different semantics of BPEL constructors. Among the different potential executable plans only one—called *executable plan*—should be selected and transformed to a workflow for effective execution.

The composition operation starts by user specification of functional and non-functional requirements and leads to an executable plan that can be handed off to runtime environment for execution. The proposed approach to support the composition operation contains three phases:

1. *Logical composition*: First, the functional requirements provided by the user are used to generate the composition graph.
2. *Physical composition*: Second, the composition graph is transformed to obtain the set of potential executable plans.
3. *Evaluation and selection*: Third, the different potential executable plans are evaluated and compared in order to select one executable plan. The latter is then transformed into a workflow and then deployed, discovered and invoked.

The service composition approach is implemented by a layered system called QoSeBroker (for QoS-enhanced Broker). The architecture of QoSeBroker is given in Figure 1. A detailed description of QoSeBroker is given in (Chakhar, 2012). At this level, we just mention that the matching operation intervenes in the logical and physical composition phases, as briefly explained in the next subsection.

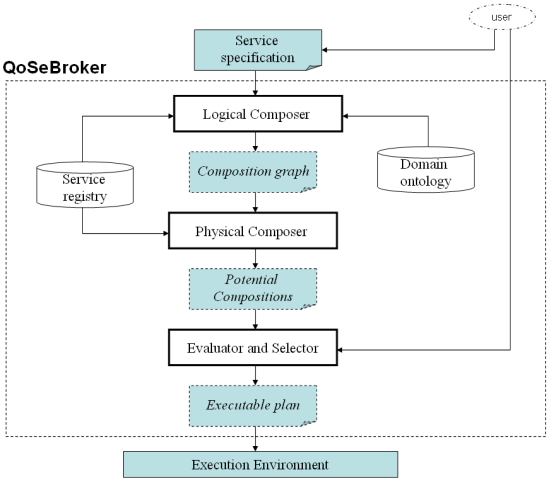


Figure 1: QoSeBroker architecture.

2.4 Service Matching Types and Process

The input for a web service composition is a set of specifications describing the capabilities of the desired service. These specifications can be decomposed into two groups (Chakhar et al., 2011): (i) functional requirements that deal with the desired functionality of the composite service, and (ii) non-functional requirements that relate to the issues like cost, performance and availability. These specifications need to be expressed in an appropriate language. In this paper, we adopt an extended version of Ontology Web Language (OWL) (Agarwal et al., 2005) for expressing functional requirements and the QoS for non-functional requirements.

Furthermore, we may distinguish three types of service matching (Chakhar, 2013):

- **Functional Attribute-level Matching:** This type of matching implies capability and property attributes and consider each matching attribute independently of the others.
- **Functional Service-level Matching:** This type of matching implies capability and property attributes but the matching operation implies attributes both independently and jointly.
- **Non-functional Matching:** This type of matching implies QoS attributes.

These different matching types intervene in different composition phases. The functional matching intervenes in the logical composition phase. Indeed, the construction of the composition graph needs first the identification of candidate service types. This operation requires matching the preconditions of a web service with the effects of another up front during filtering by using only functional requirements. The functional matching takes as input all candidate web services and outputs a set of web services that meet the user functional matching criteria. During logical composition phase, service types that fail to meet the user functional requirements are automatically eliminated. At this level, it is important to mention that functional service-level matching supports jointly attribute-level and service-level matching. Hence, the user should select either functional attribute-level matching only or functional service-level matching only.

The non-functional matching intervenes during the physical composition phase. In fact, the first step in this phase is to identify, for each service type in the composition graph, the set of the corresponding instances. For this purpose, the matching operation uses the composition graph and queries the service instances registry to associate to each service type a set of potential instances (at last one instance per type is needed). The second step is then to use non-functional attributes in order to classify web service instances into different QoS ordered classes. It is easy to see that there is no elimination of web service instances at this level.

In (Chakhar, 2013) we discussed functional attribute-level and functional service-level matching. However, the functional attribute-level matching presented in (Chakhar, 2013) assumes only conjunctive or disjunctive logical relationships between matching attributes. This paper enhances our proposal in (Chakhar, 2013) by adding generic functional attribute-level matching and non-functional QoS-oriented matching.

3 FUNCTIONAL ATTRIBUTE-LEVEL MATCHING

Functional matching may be defined as the process of discovering a service advertisement that *sufficiently* satisfies a service request (Doshi et al., 2004). Functional matching is based on the concept of *sufficiency*, which itself is based on the *similarity* measure.

3.1 Similarity Measure

A semantic match between two entities frequently involves a similarity measure. The similarity measure quantifies the semantic distance between the two entities participating in the match. As in (Doshi et al., 2004), a similarity measure is defined as follows.

Definition 5 (**Similarity Measure**). *The similarity measure, μ, of two service attributes is a mapping that measures the semantic distance between the conceptual annotations associated with the service attributes. Mathematically,*

$$\mu : A \times A \rightarrow \{\textit{Exact, Plug-in, Subsumption, Container, Part-of, Disjoint}\}$$

where A is the set of all possible attributes.

The mapping between two conceptual annotations is called:

- **Exact** map: if the two conceptual annotations are syntactically identical,
- **Plug-in** map: if the first conceptual annotation is specialized by the second,
- **Subsumption** map: if the first conceptual annotation specializes the second,
- **Container** map: if the first conceptual annotation contains the second,
- **Part-of** map: if the first conceptual annotation is a part of the second, and
- **Disjoint** map: if none of the previous cases applies.

A preferential total order is established on the above mentioned similarity maps.

Definition 6 (**Similarity Measure Preference**). *Preference amongst similarity measures is governed by the following strict total order:*

$$\textit{Exact} \succ \textit{Plug-in} \succ \textit{Subsumption} \succ \textit{Container} \succ \textit{Part-of} \succ \textit{Disjoint}$$

where $a \succ b$ means that a is preferred over b.

3.2 Computing the Similarity Measure

The computing of the similarity measure is based on the conceptual annotations associated with the requested and advertised web services. The matching process uses ontological representation of web services to infer the submission hierarchy which leads to the recognition of semantic matches despite their syntactic differences and difference in modeling abstractions between requests and advertisements.

The similarity measure, which represents the degree of match, is defined along a discrete scale in which the 'Exact' match is the highest level and is the preferable while 'Fail' is the lowest level and it represents an undesirable result. Note that the labels used to define the different similarity measures can be mapped to numerical values. However, only comparison operations are authorized (since we assumed that these values are ordinal).

Let now explain how the similarity measure μ is computed. For this purpose we consider the travel agent scenario example given in Section 2.1. Figure 2 shows a fragment of the Travel Agent Ontology. We consider an *Advertisement* for the travel agent web service who permits the client to book *Accommodation* in the specified *Destination*. Assume that the *Advertisement* has the following Inputs and Outputs:

Inputs:	*Destination*
Outputs:	*Accommodation*, *Sport*, *Cost*

We consider a *Query* from a client who planning a vacation. The client wants to make reservations for a *Hotel* and an *Excursion* at the specified *Destination*. As shown in Figure 2, the *Hotel* is a subclass of the concept *Accommodation* and *Excursion* is a subclass of the concept *Entertainment*. The Query has the following Inputs and Outputs:

Inputs:	*Destination*
Outputs:	*Hotel*, *Excursion*

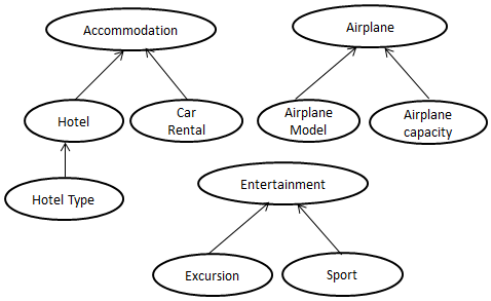


Figure 2: Travel Ontology extract.

Let now focalize on the process of matching *Query* outputs. The same reasoning applies to the process of matching the inputs. The matching process iterates over the list of output concepts of the *Query* and looks to find a match to an output concept in the *Advertisement*. Intuitively, any concept in the output represents a candidate for the match. In this particular example, the initial candidate list is $\{$*Accommodation*, *Sport*, *Cost*$\}$. The matching process first looks to see if there is a match for *Hotel*. Based on Figure 2, the match between *Hotel* and *Accommodation* will be flagged Exact. Then, the matching process looks a match for *Excursion*. Based on Figure 2, the match for *Excursion* with respect to *Accommodation*, *Sport* and *Cost* will fail (since there is no relationship between these concepts and *Excursion* concept). Hence, we can conclude that μ(*outputR*,*outputA*)='Fail' where *outputR* and *outputA* are the output attributes of requested and advertised services, respectively. Using the same reasoning, we obtain an Exact match for the input concept *Destination*, i.e., μ(*inputR*,*inputA*)='Exact' where *inputR* and *inputA* are the input attributes of requested and advertised services, respectively.

3.3 Conjunctive/Disjunctive Matching

Let S^R be the service that is requested, and S^A be the service that is advertised. A first customization of functional matching is to allow the user to specify a desired similarity measure for each attribute. A sufficient match exists between S^R and S^A in respect to a given attribute if there exists an identical attribute of S^A and the values of the attributes satisfy the desired similarity measure. A second customization of the matching process is to allow the user specifying which attributes should be utilized during the matching process, and the order in which the attributes must be considered for comparison. In order to support both customizations, we use the concept of Criteria Table, introduced by (Doshi et al., 2004), that serves as a parameter to the matching process.

Definition 7 (**Criteria Table**). *A Criteria Table, C, is a relation consisting of two attributes, C.A and C.M. C.A describes the service attribute to be compared, and C.M gives the least preferred similarity measure for that attribute. Let $C.A_i$ and $C.M_i$ denote the service attribute value and the desired measure in the ith tuple of the relation. C.N denotes the total number of tuples in C.*

Example 5. *Table 1 shows a Criteria Table example.*

A sufficient functional attribute-level conjunctive match between services is defined as follows.

Definition 8 (**Sufficient Functional Conjunctive Match**). *Let S^R be the service that is requested, and*

S^A be the service that is advertised. Let C be a criteria table. A sufficient conjunctive match exists between S^R and S^A if for every attribute in C.A there exists an identical attribute of S^R and S^A and the values of the attributes satisfy the desired similarity measure as specified in C.M. Formally,

$$\forall_i \exists_{j,k} (C.A_i = S^R.A_j = S^A.A_k) \wedge \mu(S^R.A_j, S^A.A_k) \succeq C.M_i \Rightarrow SuffFuncConjMatch(S^R, S^A) \quad 1 \leq i \leq C.N. \qquad (1)$$

Table 1: An example Criteria Table.

C.A	C.M
input	Exact
output	Exact
precondition	Subsumes
postcondition	Subsumes

The functional attribute-level conjunctive match is formalized in Algorithm 1 in (Chakhar, 2013). A less restrictive definition of sufficiency consists in using a disjunctive rule on the individual matching measures.

Definition 9 (**Sufficient Functional Disjunctive Match**). *Let S^R be the service that is requested, and S^A be the service that is advertised. Let C be a criteria table. A sufficient disjunctive match exists between S^R and S^A if for at least one attribute in C.A it exists an identical attribute of S^R and S^A and the values of the attributes satisfy the desired similarity measure as specified in C.M. Formally,*

$$\exists_{i,j,k} (C.A_i = S^R.A_j = S^A.A_k) \wedge \mu(S^R.A_j, S^A.A_k) \succeq C.M_i \Rightarrow SuffFuncDisjMatch(S^R, S^A). \qquad (2)$$

The functional attribute-level disjunctive match is formalized in Algorithm 2 in (Chakhar, 2013).

3.4 Generic Matching

In this section we extend the algorithms proposed in (Chakhar, 2013) to generic binary connectors by allowing the user to specify the conditional relationships between the capability and property attributes. First, we need to introduce the concept of sufficient single attribute match.

Definition 10 (**Sufficient Single Attribute Match**). *Let S^R be the service that is requested, and S^A be the service that is advertised. Let C be a criteria table. A sufficient match exists between S^R and S^A in respect to attribute $S^R.A_i$ if there exists an identical attribute of S^A and the values of the attributes satisfy the desired similarity measure as specified in $C.M_i$. Formally,*

$$\exists_{j,k} (C.A_i = S^R.A_j = S^A.A_k) \wedge \mu(S^R.A_j, S^A.A_k) \succeq C.M_i) \Rightarrow SuffSingleAttrMatch(S^R, S^A, A_i). \qquad (3)$$

The single attribute matching is formalized in Algorithm 1 that follows directly from Sentence (3).

Algorithm 1: SuffSingleAttrMatching.

Input : S^R, // Requested service.
S^A, // Advertised service.
C, // Criteria Table.
i, // Service attribute index.
Output: Boolean// success/fail.
while $(j \leq S^R.N)$ **do**
  if $(S^R.A_j = C.A_i)$ **then**
    Append $S^R.A_j$ to *rAttrSet*;
  end
  Assign $j \longleftarrow j+1$;
end
while $(k \leq S^A.N)$ **do**
  if $(S^A.A_k = C.A_i)$ **then**
    Append $S^A.A_k$ to *aAttrSet*;
  end
  Assign $k \longleftarrow k+1$;
end
if $(\mu(rAttrSet[i], aAttrSet[i]) \succeq C.M_i)$ **then**
  return success;
end
return fail;

The sufficient functional generic match is then defined as follows.

Definition 11 (**Sufficient Functional Generic Match**). *Let S^R be the service that is requested, and S^A be the service that is advertised. Let C be the criteria table. Let T be a complex logical clause where operands are the attributes related by logical operators (e.g. or, and, not). A sufficient functional generic match between S^R and S^A holds if and only the logical clause T holds. Formally,*

$$\textbf{\textit{Parse}}(T) \wedge \textbf{\textit{Evaluate}}(T) \Rightarrow SuffAttrGenericMatch(S^R, S^A) \qquad (4)$$

*where **Parse** and **Evaluate** are functions devoted respectively to parse and evaluate the logical expression T.*

The functional generic match is formalized in Algorithm 2, which follows directly from Sentence (4).

Example 6. *A example of a logical expression is "$T = A_5$ or (A_2 and A_3)". In this example, the matching holds when either (i) the matching in respect to attribute A_5 holds, or (ii) the matching in respect to attribute A_2 and the matching in respect to attribute A_3 hold jointly.*

3.5 Computational Complexity

Let first focalize on the complexity of Algorithm 1. The complexity of the two *while* loops in Algorithm 1 is equal to $O(S^R.N) + O(S^A.N)$. Since we generally have $S^A.N \gg S^R.N$, hence the complexity of the

two *while* loops is equal to $O(S^A.N)$. Then, the worst case complexity of Algorithm 1 is $O(S^A.N)+\alpha$ where α is the complexity of computing μ. The value of α depends on the approach used to infer μ. As underlined in (Doshi et al., 2004), inferring μ by ontological parse of pieces of information into facts and then utilizing commercial rule-based engines which use the fast Rete (Forgy, 1982) pattern-matching algorithm leads to $\alpha = O(|R||F||P|)$ where $|R|$ is the number of rules, $|F|$ is the number of facts, and $|P|$ is the average number of patterns in each rule. In this case, the worst case complexity of Algorithm 1 is $O(S^A.N)+O(|R||F||P|)$. Furthermore, we observe, as in (Doshi et al., 2004), that the process of computing μ is the most "expensive" step of the algorithms. Hence, we obtain: $O(S^A.N)+O(|R||F||P|) \asymp O(|R||F||P|)$.

The complexity of Algorithm 2 depends on the complexity of functions **Parse** and **Evaluate**. The complexity of these functions depends on the data structure used to represent the logical expression T (graph, truth tables, etc.). Clearly the complexity of **Evaluate** function is largely greater than the complexity of **Parse** function. Hence, the complexity of Algorithm 2 is $O(|R||F||P|)+O(\gamma)$ where $O(\gamma)$ is the complexity of **Evaluate** function.

Algorithm 2: SuffAttrGenericMatch.

```
Input  : S^R, // Requested service.
         S^A, // Advertised service.
         C, // Criteria table.
         T, // Logical expression.
Output: Boolean// success/fail.
if (NOT(Parse(T))) then
  |  return fail;
end
T' ⟵ T;
Z ⟵ ∅;
for (each  A_l ∈ T') do
  |  if (A_l ∉ Z) then
  |  |  t ⟵ false;
  |  |  t ⟵ SuffSingleAttrMatch(S^R, S^A, A_l);
  |  |  replace all A_l ∈ T by the value of t;
  |  |  Z ⟵ Z ∪ {A_l};
  |  end
end
if (Evaluate(T)) then
  |  return success;
end
return fail;
```

4 SERVICE-LEVEL MATCHING

The service-level matching allows the client to use two types of desired similarity: (i) desired similarity values associated with each attribute in the criteria table, and (ii) a global desired similarity that applies to the service as a whole. The service-level similarity measure quantifies the semantic distance between the requested service and the advertised service entities participating in the match by taking into account both attribute-level and service-level desired similarity measures.

Definition 12 (Sufficient Functional Service-Level Match). *Let S^R be the service that is requested, and S^A be the service that is advertised. Let C be a criteria table. Let β be the service-level desired similarity measure. A sufficient service-level match exists between S^R and S^A if (i) for every attribute in $C.A$ there exists an identical attribute of S^R and S^A and the values of the attributes satisfy the desired similarity measure as specified in $C.M$, and (ii) the value of overall similarity measure satisfies the desired overall similarity measure β. Mathematically,*

$$[\forall i \quad (SuffSingleAttrMatch(S^R, S^A, A_i)) \quad 1 \le i \le C.N] \wedge [\exists j_1, \cdots, j_i, \cdots, j_N \quad (\zeta(s_{1,j_1}, \cdots, s_{i,j_i}, \cdots, s_{N,j_N}) \succeq \beta)] \Rightarrow SuffFuncServiceLevelMatch(S^R, S^A) \qquad (5)$$

where ζ is an aggregation rule; and for $i = 1, \cdots, N$ and $j_i \in \{j_1, \cdots, j_N\}$:

$$s_{i,j_i} = \mu(S^R.A_i, S^A.A_{j_i}).$$

The parameter β may be any of the maps given in Section 3.1. The functional service-level matching in formalized Algorithm 4 in (Chakhar, 2013). The aggregation rule ζ used in the definition above is a tool to combine the similarity measures into a single similarly measure. In (Chakhar, 2013), we defined ζ as follows:

$$\zeta : F_1 \times \cdots \times F_N \rightarrow \{\text{Exact, Plug-in, Subsumption, Container, Part-of, Disjoint}\}$$

where F_j=$\{$Exact, Plug-in, Subsumption,Container, Part-of, Disjoint$\}$ ($j = 1, \cdots, N$); and N is the number of attributes included in the criteria table. The similarity maps and the corresponding strict total order given in Section 3.1 still apply here. Since the similarity measures are defined on an ordinal scale, there are only a few possible aggregation rules that can be used to combine similarity measures (Chakhar, 2013): Minimum, Maximum, Median, Floor and Ceil. The Floor and Ceil rules apply only when there is an even number of similarity measures (which leads to two median values).

5 QoS-ORIENTED MATCHING

The QoS-oriented matching concerns QoS attributes only and applies to service instances that verify functional requirements. The objective of QoS matching

is to assign to each instance an overall QoS level. Instead of sorting services from best to worst, we propose to categorize them into an ordered set of QoS classes $\mathbf{Cl} = \{Cl_1, \cdots, Cl_p\}$, such that the higher the class, the higher the QoS level.

The computing of overall QoS level for each instance requires the use of a multicriteria aggregation rule. In this paper, we distinguish two types of aggregation rules: (i) simple majority, and (ii) simple majority with veto.

5.1 Simple Majority-based Matching

To formalize the first version of QoS matching, we need to introduce some new concepts.

Definition 13 (QoS Attribute Table). *A QoS Attribute Table, Q, is a relation consisting of three attributes, Q.A, Q.T and Q.S. Q.A describes the service attribute to be compared, Q.T gives the attribute type and Q.S specifies the scale type. Two types of attributes are distinguished: gain and cost. The gain attributes are those to be maximized while cost attributes are those to be minimized. The scale may be nominal, ordinal, cardinal or ratio. Let $Q.A_i$, $Q.T_i$ and $Q.S_i$ denote the service attribute value, the attribute type and the scale type of the ith tuple of the relation. Let Q.N be the total number of tuples in Q.*

Example 7. *Table 2 shows a QoS Attribute Table example. It specifies the parameters of four QoS attributes: response time (A_1), availability (A_2), security (A_3) and cost (A_4).*

Table 2: An example QoS Attribute Table.

$Q.A$	$Q.T$	$Q.S$
A_1: Response time	cost	Cardinal
A_2: Availability	gain	Cardinal
A_3: Security	gain	Ordinal
A_4: Cost	cost	Cardinal

Definition 14 (Boundary Matrix). *A Boundary Matrix, B, consisting of a pairwise matrix composed of $p-1$ columns $B_1, \cdots, B_{p-1}$ and N rows corresponding to the number of QoS attributes.*

Example 8. *An example of Boundary Matrix is given in Table 3. It specifies three boundaries in respect to the QoS given in Table 2. Table 3 defines four QoS classes.*

The attribute type and scale parameters should be used to control input data, especially the definition of boundaries.

Definition 15 (Weight Table). *A Weight Table, W, is a relation consisting of two attributes, W.A and W.V. W.A describes the service attribute and W.V specifies the weight of this attribute. Let $W.A_i$ and $W.V_i$ denote the service attribute and the attribute weight value in the ith tuple of relation W. The weights values must sum to 1.*

Table 3: An example Boundary Matrix.

$Q.A_i$	B_1	B_2	B_3
Response time	11	9.25	8
Availability	0.2	0.3	0.51
Security	2	3	4
Cost	4	3.5	3

Example 9. *An example of Weight Table is given in Table 4.*

Table 4: An example Weight Table.

$W.A$	$W.V$
Response time	0.325
Availability	0.325
Security	0.175
Cost	0.175

Definition 16 (Concordance Power). *Let $h \in \{1, \cdots, p\}$. The concordance power for the outranking of advertised service S^A over boundary B_h is computed as follows:*

$$\Phi(S^A, B_h) = \sum_{i \in L_1(S^A, h)} W.V_i$$

where: $L_1(S^A, B_h) = \{i : S^A.A_i \succeq B_h.A_i \wedge Q.T_i = \text{'gain'}\} \cup \{i : S^A.A_i \preceq B_h.A_i \wedge Q.T_i = \text{'cost'}\} \cup \{i : S^A.A_i = B_h.A_i \wedge Q.S_i = \text{'nominal'}\}$.

Example 10. *Let consider the service instances given in Table 5. Based on the definition above we obtain $L_1(s_8, B_1) = \{1,3,2,4\}$, $L_1(s_8, B_2) = \{2,3,4\}$ and $L_1(s_8, B_3) = \{4\}$. This leads to: $\Phi(s_8, B_1) = 1$, $\Phi(s_8, B_2) = 0.675$ and $\Phi(s_8, B_3) = 0.175$.*

Table 5: Web service instances.

s_i	A_1	A_2	A_3	A_4
s_8	12.82	0.34296	3	2.74
s_9	10.92	0.15	1	2.08
s_{10}	9.52	0.51	4	2.5

Definition 17 (Sufficient Simple Majority-based QoS Matching). *Let S^R be the service that is requested and S^A be the service that is advertised. Let $S^R.Q$ be the list of QoS attributes to be utilized for matching. Service S^A is assigned to QoS class Cl_h if for every QoS attribute of S^R there is exists an identical attribute of S^A and the value of the Concordance*

Power is greater or equal to the credibility threshold λ *and* $\Phi(S^A, B_h) \geq \Phi(S^A, B_{h'})$ *for every* $h < h'$. *Formally,*

$$Argmax_h[\exists_{j,k}(Q.A_i = S^R.A_j = S^A.A_k) \wedge \Phi(S^A, B_h) \succeq \lambda] \Rightarrow QoSMajorityMatch(S^R, S^A, Cl_h). \quad (6)$$

where $\lambda \in [0.5, 1]$.

The parameter λ, called *credibility threshold*, ensures that at least 50% of attributes are in favor of the assignment of the advertised service to the considered QoS class. Hence, a service S^A is assigned to class Cl_h if and only if there is a "sufficient" majority of attributes in favor of assigning S^A to Cl_h.

This first version of QoS matching is formalized in Algorithm 3. This algorithm compares S^A to each of the boundaries staring from the highest one and stops once a sufficient QoS measure holds. The function *ConcordancePower*, which computes the concordance power as in Definition (16), is given in Algorithm 4.

Algorithm 3: QoSMajorityMatching.

```
Input  : S^A, // Advertised service.
         λ, // Credibility threshold.
         Q, // QoS Table.
         B, // Boundary Matrix.
         W, // Weight Table.
Output: Cl = {Cl_1, ··· , Cl_p}// Global QoS classes.
p ⟵ number of QoS classes;
Cl_i ⟵ ∅, ∀i = 1, ··· , p;
U' ⟵ instances of S^A;
for (all u ∈ U') do
    h ⟵ p − 1;
    assigned ⟵ False;
    while (h ≥ 0 ∧ NOT(assigned)) do
        if (ConcordancePower(u, h) ≥ λ) then
            Cl_{h+1} ⟵ Cl_{h+1} ∪ {u};
            assigned ⟵ true;
        end
        h ⟵ h − 1;
    end
end
Cl ⟵ {Cl_1, ··· , Cl_p};
return Cl;
```

Example 11. *Let* $\lambda = 0.65$. *Based on the previous example, we conclude that* s_8 *in Table 5 is assigned by Algorithm 3 to QoS class level 3 since* $\Phi(s_2, B_3) = 0.175 < \lambda$ *and* $\Phi(s_8, B_2) = 0.675 > \lambda$.

5.2 Simple Majority and Veto Based QoS Matching

A further customization consisting in the support of the veto effect by taking into account the opposition of

Algorithm 4: ConcordancePower.

```
Input  : S^A, // Advertised service.
         h, // Integer (Boundary index).
Output: Φ(S^A, B_h)// Concordance Power.
s ⟵ 0;
for (all Q.A_i ∈ Q.A) do
    switch Q.T_i do
        case ('gain')
            if (S^A.A_i ≥ B_h.A_i) then
                s ⟵ s + W.V_i;
            end
        end
        case ('cost')
            if (S^A.A_i ≤ B_h.A_i) then
                s ⟵ s + W.V_i;
            end
        end
        case ('nominal')
            if (S^A.A_i = B_h.A_i) then
                s ⟵ s + W.V_i;
            end
        end
    end
end
return s;
```

minority attributes, i.e., attributes that do not belong to set L_1. The definition of QoS matching with support of veto effect requires the introduction of some new concepts.

Definition 18 (Discordance Power). *Let* $h \in \{1, \cdots, p\}$. *The discordance power for the outranking of advertised service* S^A *over boundary* B_h *is computed as follows:*

$$\Psi(S^A, B_h) \quad = \quad \prod_{k=1}^{k=N} Z_k(S^A.A_k, B_h.A_k)$$

where:

$$Z_k(S^A.A_k, B_h.A_k) \quad = \quad \begin{cases} \frac{1-W.V_k}{1-\Phi(S^A,B_h)}, & if\ W.A_k > \Phi(S^A, B_h)) \\ & \wedge \quad k \in L_2(S^A, B_h) \\ 1, & otherwise. \end{cases}$$

with $L_2(S^A, B_h) = \{i : S^A.A_i \prec B_h.A_i \wedge Q.T_i = 'gain'\} \cup \{i : S^A.A_i \succ B_h.A_i \wedge Q.T_i = 'cost'\} \cup \{i : S^A.A_i \neq B_h.A_i \wedge Q.S_i = 'nominal'\}$.

Example 12. *Let consider the service instances given in Table 5. Based on the definition above we obtain* $L_2(s_8, B_1) = \{1\}$, $L_2(s_8, B_2) = \{1\}$ *and* $L_2(s_8, B_3) = \{1, 2\}$. *This leads to:* $\Psi(s_8, B_1) = 0.818$, $\Psi(s_8, B_2) = 0.818$ *and* $\Psi(s_2, B_3) = 0.670$.

Definition 19 (Credibility Index). *Let* $h \in \{1, \cdots, p\}$. *The credibility index for the outranking of advertised service* S^A *over boundary* B_h *is computed as follows:*

$$\sigma(S^A, B_h) \quad = \quad \Phi(S^A, B_h) \cdot \Psi(S^A, B_h).$$

Example 13. *Based on Examples 10 and 12, we obtain* $\sigma(s_8,B_1) = 0.818$, $\sigma(s_8,B_2) = 0.552$ *and* $\sigma(s_8,B_3) = 0.117$.

Definition 20 (Sufficient Majority With Veto Based QoS Matching). *Let* S^R *be the service that is requested and* S^A *be the service that is advertised. Let* $S^R.Q$ *be the list of QoS attributes to be utilized for matching. Service* S^A *is assigned to QoS class* Cl_h *if for every QoS attribute of* S^R *there is exists an identical attribute of* S^A *and the value of the Credibility index is greater or equal to the credibility threshold* λ *and* $\sigma(S^A,B_h) \geq \sigma(S^A,B_{h'})$ *for every* $h < h'$. *Formally,*

$$Argmax_h[\exists_{j,k}(Q.A_i = S^R.A_j = S^A.A_k) \wedge \sigma(S^A,B_h) \geq \lambda] \Rightarrow QoSMajorityVetoMatching(S^R,S^A,Cl_h). \quad (7)$$

where $\lambda \in [0.5,1]$.

According to this definition, a service S^A is assigned to class Cl_h if and only if: (i) there is a "sufficient" majority of attributes in favor of assigning S^A to Cl_h, and (ii) when the first condition holds, none of the minority of attributes shows an "important" opposition to the assignment of S^A to Cl_h.

The algorithm for the second version of QoS matching is similar to Algorithm 3. We simply need to replace the test "$ConcordancePower(u,h) \geq \lambda$" by "$CredibilityIndex(u,h) \geq \lambda$". The function *CredibilityIndex*, which computes the credibility index as in Definition (19), is given in Algorithm 5.

Algorithm 5: CredibilityIndex.

```
Input  : S^A, // Advertised service.
         h, // Integer (Boundary index).
Output: σ(S^A,B_h)// Credibility Index.
z ⟵ 1;
s ⟵ ConcordancePower(S^A,h);
for (all Q.A_i ∈ Q.A) do
    switch Q.T_i do
        case ('gain')
            if (S^A.A_i < B_h.A_i ∧ W.A_i > s) then
                z ⟵ z · (1−W.A_i)/(1−s);
            end
        end
        case ('cost')
            if (S^A.A_i > B_h.A_i ∧ W.A_i > s) then
                z ⟵ z · (1−W.A_i)/(1−s);
            end
        end
        case ('nominal')
            if (S^A.A_i ≠ B_h.A_i ∧ W.A_i > s) then
                z ⟵ z · (1−W.A_i)/(1−s);
            end
        end
    end
end
return s · z;
```

Example 14. *Let* $\lambda = 0.65$. *Based on the data and results of the previous example, we conclude that* s_8 *in Table 5 is assigned by majority with veto Algorithm to QoS class level 2 since* $\sigma(s_8,B_1) = 0.818 > \lambda$, *and* $\sigma(s_8,B_2) = 0.552 < \lambda$ *and* $\sigma(s_8,B_3) = 0.117 < \lambda$.

The previous example shows that the QoS of instance s_8 has decreased (from 3 to 2) in comparison to Example (11). This because the weights of attributes which are against the assignment of instance s_8 to QoS class 3 have been taken into account.

5.3 Computational Complexity

Algorithm 3 runs in $O(mp \times |U|)$, where m is the number of QoS attributes and p is the number of QoS classes. The second version of Algorithm 3 based on simple majority with veto and which is not provided in this paper, runs in $O(2mp \times |U|)$. Note that Algorithm 4 runs in $O(m)$ and Algorithm 5 runs in $O(2m)$.

6 ILLUSTRATION

Consider again the example given in Section 3.2. Assume that the Criteria Table is as in Table 6. Based on the discussion given in Section 3.2, the desired similarity specified by this table holds only for *Destination* and *Hotel*. Hence, the attribute-level conjunctive matching algorithm will fail but the attribute-level disjunctive matching algorithm will success.

Table 6: Criteria Table.

C.A	*C.M*
input	Exact
output	Exact

Let now assume that the user has specified the following logical expression:

expression:	*Destination* and (*Hotel* or *Excursion*)

The desired similarity specified by the Criteria Table holds only for *Destination* and *Hotel*. But since the expression "*Destination* and (*Hotel* or *Excursion*)" will be flagged true, the functional generic matching algorithm will success.

Let now focalize on service-level functional matching. Suppose that the *Advertisement* has the following Inputs, Outputs and Categories:

Inputs:	*Destination*
Outputs:	*Accommodation*, *Entertainment*, *Cost*
Categories:	*AirplaneModel*

Suppose also that *Query* has the following Inputs, Outputs and Categories:

Inputs:	*Destination*
Outputs:	*HotelType*, *Excursion*
Categories:	*AirplaneModel*

Based on the same reasoning as earlier, the match for *Destination*, *Excursion* and *AirplaneModel* concepts will be flagged Exact; and the match for concept *HotelType* will be flagged Plug-in. Assume that the Criteria Table is as in Table 7.

Table 7: Criteria Table.

C.A	*C.M*
input	Exact
output	Exact
service category	Subsumption

The instantiation of the aggregation rule ζ (see Section 4) will then look like ζ(Exact, min{Plug-in, Exact}, Exact) which leads to ζ(Exact, Plug-in, Exact). Then, the result of aggregation according to different aggregation rules (except Median which does not apply here) is given in Table 8.

Table 8: Result of aggregation.

Aggregation rule(ζ)	Minimum	Maximum	Floor	Ceil
Result	Plugin	Exact	Plugin	Exact

Let now focalize on non-functional matching and consider the list of potential compositions given in Table 9. We assume that these compositions have meet the functional requirements of the user. Table 9 shows the evaluation of the compositions in respect to four QoS attributes (i.e. response time (A_1), availability (A_2), security (A_3), and cost (A_4) attributes) given in Table 2. The objective is to classify the compositions into different ordered categories. For the purpose of this example, we assume that the four categories defined by Table 3 and the weights given in Table 4 have been used.

The final classifications obtained by the simple majority and simple majority with veto algorithms where the credibility threshold is $\lambda = 0.65$ are given in Table 9. In this table, we can see that both simple majority and majority with veto algorithms (denoted Algo 3 and Algo 3' in Table 9, respectively) assign instances s_3 and s_{10} to the best QoS class. We remark also that both algorithms assign instances s_5 and s_9 to the worst QoS class. Which is interesting to see in Table 9 is the role of veto effect in the assignment of instances s_8 and s_{13}. Indeed, the QoS of both instances have been decreased (from 3 to 2 for instance s_8 and from 2 to 1 for instance s_{13}) by the majority with veto algorithm. This happens because the weights of attributes which are against the assignment—of instance s_8 to QoS class 3 and instance s_{13} to QoS class 2—have been taken into account.

Table 9: Potential compositions and final classification for λ=0.65.

s_i	$A_1(s_i)$	$A_2(s_i)$	$A_3(s_i)$	$A_4(s_i)$	Algo 3	Algo 3'
s_1	9.2	0.45946	1	2.48	3	3
s_2	8.12	0.41817	1	2.68	3	3
s_3	8	0.53	4	2.78	4	4
s_4	8.19	0.46967	2	3.24	3	3
s_5	11.15	0.19	1	2.74	1	1
s_6	7.42	0.40317	2	3.38	3	3
s_7	7.72	0.36676	2	3.18	3	3
s_8	12.82	0.34296	3	2.74	3	2
s_9	10.92	0.15	1	2.08	1	1
s_{10}	9.52	0.51	4	2.5	4	4
s_{11}	10.12	0.53294	3	2.68	3	3
s_{12}	10.42	0.48356	1	2.32	2	2
s_{13}	12.52	0.2	1	3.14	2	1
s_{14}	8.42	0.48	1	2.82	3	3
s_{15}	10.32	0.48	4	2.16	3	3

7 EXPERIMENTAL RESULTS

For the purpose of experimental tests, we adopted the same architecture proposed by (Bellur and Kulkarni, 2007). We used the Protege editor Protege (http://protege.stanford.edu/) to browse and edit OWL ontologies. The Mindswap OWL-S API has been used to load the OWL ontologies into the Knowledge Base and parse the OWL-S Queries and Advertisements. We used the Pellet reasoner Pellet (http://pellet.owldl.com/) to classify the loaded ontologies and Jena API JENA (http://jena.sourceforge.net) to query the reasoner for concept relationships. To test the algorithms, we used 14 ontologies from the OWTLS-TC (service retrieval test collection from SemWebCentral) in our Knowledge Base. About 700 advertisements from OWLS-TC were loaded into the advertisement repository. All measurement points shown are average results taken from 30 runs. The data sets for clients and providers were randomly generated.

The result of experimental tests is given in Figures 3 and 4. Figure 3 shows the execution time of Algorithms 1, 2 and 3. Two versions (denoted Algo 3 and Algo 3' in this figure) of Algorithm 3 have been tested: the first version (Algo 3) uses the simple majority and the second version (Algo 3') uses simple majority with veto. Figure 4 shows the evolution of execution time in respect to the number of instances for the two versions of Algorithm 3. As shown in this figure, it is easy to see that the execution time varies

asymptotically for both versions of Algorithm 3. We note, however, that the second version (Algo 3') is more expensive in execution time. This because the second version of Algorithm 3 needs to execute Algorithms 4 and 5 (for computing the concordance and discordance powers) while the first version (Algo 3) of Algorithm 3 needs to execute only Algorithm 4.

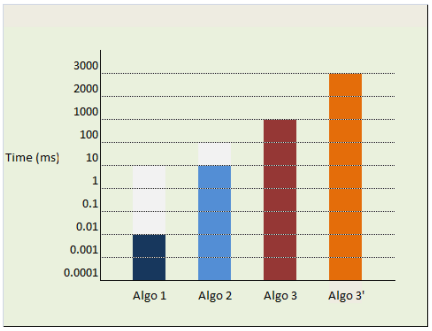


Figure 3: Execution time of algorithms.

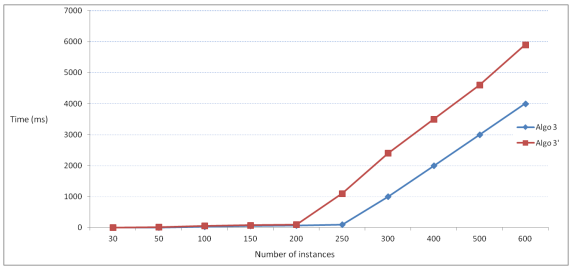


Figure 4: Execution time vs number of instances.

8 RELATED WORK

In this section we discuss some matchmaking frameworks in respect to several characteristics. The first characteristic is related to the *support of customization* which is an important issue in practice, as recognized by (Doshi et al., 2004). Most of proposed matching systems ignore this point and only a few ones take into account this aspect. (Doshi et al., 2004), for instance, present a parameterized semantic matchmaking framework that enables the user to specify the matched attributes and the order in which attributes are compared. In (Doshi et al., 2004), the sufficiency condition defined by the authors is very strict. This problem has been addressed in (Chakhar, 2013) and in this paper by relaxing matchmaking conditions and supporting three types of matching.

The second characteristic concerns the *type of attributes used in the matching operation*. Most of existing matchmaking frameworks (Ben Mokhtar et al., 2006; Guo et al., 2005; Li and Horrocks, 2003) use only service capability as the criteria for the match. The authors in (Doshi et al., 2004) distinguish two types of matching attributes: capability and property. (Ludwig, 2011) proposes two approaches to service selection based on QoS attributes. (Sathya et al., 2011) discuss various techniques of QoS based service selection. (Krithiga, 2012) proposes a QoS-based web service selection based on a stochastic optimization. (Xia et al., 2011) propose a QoS-aware web service selection algorithm based on clustering. The framework presented in this paper identify three types of matching attributes (capability, property, and QoS) by subdividing property attributes set into two sets of attributes: those directly related to the QoS and those which are not. We think that the proposed framework enhances the above cited proposals, especially the work of (Doshi et al., 2004).

The third characteristic is related to the *method used to compare the requested and advertised services*. Most of existing proposals use simple syntactic and strict capability-based search. (Doshi et al., 2004) present a semantic matchmaking framework that avoids the limitations of strict capability-based matchmaking. (Fu et al., 2009) transform the problem of matching web services to the computation of semantic similarity between concepts in domain ontology using a semantic distance measure. (Bellur and Kulkarni, 2007) improve (Paolucci et al., 2002)'s matchmaking algorithm and propose a greedy-based algorithm that relies on the concept of matching bipartite graphs. In this paper we adopted and extended the semantic matchmaking framework proposed by (Doshi et al., 2004).

The fourth characteristic concerns the *support of the multicriteria evaluation*. There are a few proposals that explicitly support multicriteria evaluation, e.g. (Cui et al., 2011; Jeong et al., 2007; Menascé, 2004; Menascé and Dubey, 2007; Zeng et al., 2003). Most of them use weighted-sum like aggregation techniques. (Zeng et al., 2003) use linear programming techniques to compute the optimal execution plans for web service. (Menascé, 2004) considers two evaluation criteria (time and cost) and assigns to each one a weigh. The best composition of Web services is then decided on the basis of the optimum combined score. (Menascé and Dubey, 2007) propose a service selection QoS broker by maximizing a utility function. We note, however, that this type of methods have two main shortcomings: (i) they accept only numerical data and (ii) may lead to the compensation problem since low values may be counterbalanced by high values. The approach used in this paper accepts any type of data and resolves the compensation problem.

9 CONCLUSION

We presented a QoS-aware semantic matching framework. The framework supports three types of matching: functional attribute-level matching, functional service-level matching, and QoS-based matching. A series of highly customizable algorithms are advertised for each type of matching.

Several issues need to be further investigated. First, the reduction of the number of parameters required from the user by automatically generating the boundaries of QoS classes. Second, the use of the rough sets theory-based classification (Greco et al., 2001) for assigning instances to QoS classes. Third, the use of multicriteria ranking/choice approaches instead of the classification approach used in this paper.

REFERENCES

Agarwal, V., Chafle, G., Dasgupta, K., Karnik, N., Kumar, A., Mittal, S., and Srivastava, B. (2005). Synthy: A system for end to end composition of web services. *Journal of Web Semantics*, 3:311–339.

Bellur, U. and Kulkarni, R. (2007). Improved matchmaking algorithm for semantic web services based on bipartite graph matching. In *IEEE International Conference on Web Services*, pages 86–93, Salt Lake City, Utah, USA.

Ben Mokhtar, S., Kaul, A., Georgantas, N., and Issarny, V. (2006). Efficient semantic service discovery in pervasive computing environments. In *ACM/IFIP/USENIX 2006 International Conference on Middleware*, pages 240–259, Melbourne, Australia.

Chakhar, S. (2012). QoS-enhanced broker for composite web service selection. In *Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS 2012)*, pages 533–540, Sorrento-Naples, Italy.

Chakhar, S. (2013). Parameterized attribute and service levels semantic matchmaking framework for service composition. In *Fifth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2013)*, pages 159–165, Seville, Spain.

Chakhar, S., Youcef, S., Mousseau, V., Mokdad, L., and Haddad, S. (2011). Multicriteria evaluation-based conceptual framework for composite web service selection. Working Paper, Lamsade, University Paris-Dauphine, France. http://basepub.dauphine.fr/ xmlui/bitstream/handle/123456789/5283/multicriteria mokdad.PDF?sequence=2.

Cui, L., Kumara, S., and Lee, D. (2011). Scenario analysis of web service composition based on multi-criteria mathematical goal programming. *Service Science*, 3(4):280–303.

Doshi, P., Goodwin, R., Akkiraju, R., and Roeder, S. (2004). Parameterized semantic matchmaking for workflow composition. IBM Research Report RC23133, IBM Research Division.

Forgy, C. (1982). Rete: A fast algorithm for the many patterns/many objects match problem. *Artificial Intelligence*, 19(1):17–37.

Fu, P., Liu, S., Yang, H., and Gu, L. (2009). Matching algorithm of web services based on semantic distance. In *International Workshop on Information Security and Application (IWISA 2009)*, pages 465–468, Qingdao, China.

Greco, S., Matarazzo, B., Slowinski, R., and Stefanowski, J. (2001). An algorithm for induction of decision rules consistent with the dominance principle. In Ziarko, W. and Yao, Y., editors, *Rough Sets and Current Trends in Computing*, volume 2005 of *LNCS*, pages 304–313. Springer Berlin Heidelberg.

Guo, R., Le, J., and Xiao, X. (2005). Capability matching of web services based on OWL-S. In *Sixteenth International Workshop on Database and Expert Systems Applications*, pages 653–657.

Hao, H., Haas, H., and Orchard, D. (2004). Web services architecture usage scenarios. W3C Working Group Note 11, IBM Research Division.

Jeong, B., Cho, H., Kulvatunyou, B., and Jones, A. (2007). A multi-criteria web services composition problem. In *IEEE International Conference on Information Reuse and Integration(IRI 2007)*, pages 379–384.

Krithiga, R. (2012). QoS-aware web service selection using SOMA. *Global Journal of Computer Science and Technology*, 12(10):46–51.

Li, L. and Horrocks, I. (2003). A software framework for matchmaking based on semantic web technology. In *12th International World Wide Web Conference*, pages 331–339, Budapest, Hungary.

Ludwig, A. (2011). Memetic algorithm for web service selection. In *Third Workshop on Biologically Inspired Algorithms for Distributed Systems*, BADS '11, pages 1–8, New York, NY, USA. ACM.

Menascé, D. (2004). Composing web services: A QoS view. *IEEE Internet Computing*, 8(6):88–90.

Menascé, D. and Dubey, V. (2007). Utility-based QoS brokering in service oriented architectures. In *IEEE International Conference on Web Services (ICWS 2007)*, pages 422–430.

Paolucci, M., Kawamura, T., Payne, T., and Sycara, K. (2002). Semantic matching of web services capabilities. In *First International Semantic Web Conference on The Semantic Web*, pages 333–347, Sardinia, Italy.

Sathya, M., Swarnamugi, M., Dhavachelvan, P., and Sureshkumar, G. (2011). Evaluation of QoS based web-service selection techniques for service composition. *International Journal of Software Engineering*, 1(5):73–90.

Xia, Y., Chen, P., Bao, L., Wang, M., and Yang, J. (2011). A QoS-aware web service selection algorithm based on clustering. In *IEEE International Conference on Web Services (ICWS)*, pages 428–435.

Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., and Sheng, Q. (2003). Quality driven web services composition. In *12th International Conference on World Wide Web*, pages 411–421, New York, NY, USA. ACM.

Low Complexity Spatial Similarity Measure of GPS Trajectories

Radu Mariescu-Istodor, Andrei Tabarcea, Rahim Saeidi and Pasi Fränti
Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Joensuu, Finland
{*radum, tabarcea, rahim, franti*}*@cs.uef.fi*

Keywords: GPS Trajectory, Spatial Similarity, MGRS, Cell Approximation, Sampling Frequency, Interpolation, Dilation.

Abstract: We attack the problem of trajectory similarity by approximating the trajectories using a geographical grid based on the MGRS 2D coordinate system. We propose a spatial similarity measure which is computationally feasible for big data collections. The proposed measure is based on cell matching with a similarity metric drawn from Jaccard index. We equip the proposed method with interpolation and dilation to overcome the problems missing data and different sampling frequencies when comparing two trajectories. The proposed measure is implemented online in the framework of Mopsi[a].

[a] cs.uef.fi/mopsi

1 INTRODUCTION

In recent years, GPS technology has been widely available in consumer devices, especially in smartphones[1], which count as more than a half on total mobile phone sales[2]. Furthermore, most of the users utilize their phone to find their location, amongst other services[3]. The wide availability of GPS-enabled smartphones that are also connected to the Internet has made the collection of large amount of location-based data possible. Such data includes geo-tagged photos, videos and geographical trajectories. Collecting geographical trajectories has practical applications in fleet management, sports tracking, recommending tourist trajectories, improving navigation and determining mobility patterns.

Having a large-scale collection of GPS trajectories raises the challenge of how to organize the data, how to present it in a meaningful way and how to filter out irrelevant data. Computing *trajectory similarity* is a tool that can be used in addressing those challenges (Agrawal et al., 1993). A problem in computing similarity of GPS trajectories is that the large amount of data does not permit processing raw trajectories in real time.

Time series analysis of one-dimensional data across the time has been used for analyzing stock changes, weather data and biomedical measurements (Hamilton, 1994; Chan and Fu, 1999; Worsley and Friston, 1995; Lange and Naumann, 2011). Despite the significant research output on time series analysis, the concept of computing similarity for traces of moving objects in the framework of spatio-temporal databases has been studied much less. Finding *k-nearest* trajectories, indexing and clustering of spatio-temporal data are among the recent directions of research with many applications to make queries in moving object databases (Frentzos et al., 2007a; Ni and Ravishankar, 2007; Frentzos et al., 2007b; Güting et al., 2010; Pelekis et al., 2011). These algorithms can be applied also for measuring the trajectory similarity (Hu and Steenkiste, 2006).

Using *Euclidean distance* is not practical for the case that the length of two trajectories are not equal (Yanagisawa et al., 2003). *Dynamic time warping* handles matching two sequences of different length but it is very sensitive to noisy data (Berndt and Clifford, 1994). Algorithms like *longest common subsequence* (LCS) (Vlachos et al., 2002b; Vlachos et al., 2002a) or *edit distance on real sequence* (EDR) (Chen et al., 2005) are designed to account for noisy and missing data but they are not perturbation free. Considering M trajectories of N points on average, the computational complexity of these algorithms is at minimum $O(M^2 \cdot N^2)$. Hence, these algorithms cannot provide real-time results when dealing with a large collection of data.

These algorithms do not utilize time stamps. By

[1] abiresearch.com/research/product/1005746-mobile-device-user-interfaces

[2] gartner.com/newsroom/id/2623415

[3] pewinternet.org/Reports/2012/Location-based-services.aspx

using the timing information a complete movement profile can be provided and the similarity of two trajectories can be used in trajectory clustering applications. The similarity measurement in LCS and EDR are based on point-to-point distance calculations. In the event of having two trajectories with different sampling frequency, LCS and EDR cannot provide correct similarity measure (Frentzos et al., 2007b). Although it is always possible to use a trajectory reduction or approximation algorithm to represent a trajectory with far less representatives for similarity calculation, the quality of such an approximation algorithm and overhead computational complexity is debatable (Ni and Ravishankar, 2007).

In this paper, we propose a fast method of computing trajectory similarity by approximating the trajectories using a geographical grid based on a 2D coordinate system. This process reduces a trajectory from points to cells with order of magnitude less details in representation and subsequently in distance calculations. We employ an asymmetric similarity metric inspired by Jaccard index. Dealing with GPS data collection, it is common to have bunch of data points lost or compare trajectories traveled by car with walking speed trajectories. We propose interpolation and dilation of trajectories represented as cells to overcome these difficulties. In the results section we simulate missing data and trajectory sampling frequency mismatch with two example trajectories and demonstrate the efficiency of the proposed approach. Conclusions are drawn after the discussion of results.

2 MOPSI

Mopsi is a research project location-based service developed at the University of Eastern Finland by Speech and Image Processing Group from the School of Computing. (Fränti et al., 2011) Mopsi offers multiple applications of location-aware systems, being a test-bed for various research topics that involve location-aware data. It contains tools for collecting, processing and displaying location-based data, such as photos or trajectories, along with social media integration. The main topics addressed in Mopsi are collecting location-based data, mining location data from web pages, processing, storing and compressing of GPS trajectories, detecting transportation mode from GPS trajectories, recommending *points of interest*, using location information in social networks, detecting users actions by using their location and building location-based games with the help of user-generated collections.

Location-based data is very common among web-

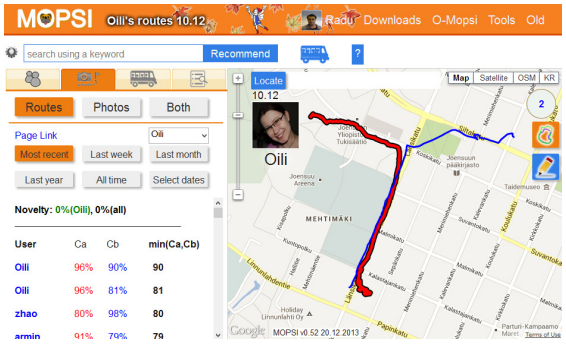


Figure 1: Mopsi application on web showing an example of two trajectories which display a common region.

pages, especially when their content describe commercial services, landmarks or public institutions. However, the location data is more commonly presented in a human-readable way and not as geographical coordinates, which are more accurate and easier to be automatically identified. We propose a method to automatically identify location information from web-pages by detecting postal addresses (Fränti et al., 2010).

Mopsi provides tools to collect GPS trajectories and it includes more than 9000 trajectories composed of over 7 million points by the end of 2013. Mopsi uses fast retrieval and displaying of the data (Waga et al., 2013) based on GPS trajectory polygonal approximation (Chen et al., 2012a). GPS trajectories are also compressed for optimizing storage space (Chen et al., 2012b). Transport mode information can be also retrieved by automatically analyzing GPS trajectories (Waga et al., 2012). The algorithm uses a second order Markov model to segment the trajectories and to detect car, bicycle, running or walking transportation modes.

The relevance of location-based media can be assessed by considering several aspects such as time, location, content or social network (Fränti et al., 2011), which are used to create a context for each user. A personalized recommender system can recommend relevant data based on user location and user context (Waga et al., 2011). Such data can be geotagged photos, services confirmed by administrators or GPS trajectories. Users can share their location in real-time by using mobile phone location-aware applications. This allows for the detection of various location-based actions such as meetings, visiting or passing-by *points of interest* (Mariescu-Istodor, 2013). Mopsi also includes location-based games, such as O-Mopsi (Tabarcea et al., 2013), which is an orienteering game using the data from a user-generated photo collection.

Mopsi provides tools for collecting location-based

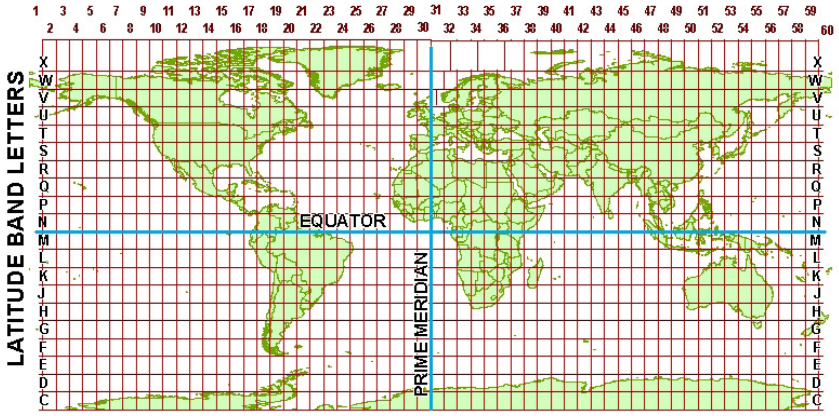


Figure 2: MGRS grid zones (source[4]).

data with mobile devices. It is available on most mobile operating systems (Android, iOS, Windows Phone, Symbian). The server-side processes the data collected by the user and displays the data collection. It also provides social features and integration which social media, with functionalities such as chatting, friends tracking and sharing data to Facebook. The Mopsi routes module provides tools for trajectory recording and displaying the large amount of data in reasonable time. Trajectory similarity is the newest addition to the Mopsi routes module.

3 TRAJECTORIES

In Mopsi we record a user's location at a certain time as a point $\mathbf{p}_k = (x_k, y_k, t_k)$, where x_k is the latitude, y_k is the longitude and t_k is the timestamp of point k. An ordered sequence of these points, defines a spatial trajectory $\mathbf{R} = (\mathbf{p}_1, \ldots, \mathbf{p}_K)$. We calculate the similarity between a reference trajectory $\mathbf{R}_a$ and all the other $M-1$ trajectories in the database, $\mathbf{R}_m$, $m = 1, \ldots, M$.

The similarity of two trajectories can be calculated as the Jaccard index:

$$J(\mathbf{R}_a, \mathbf{R}_m) = \frac{|\mathbf{R}_a \cap \mathbf{R}_m|}{|\mathbf{R}_a \cup \mathbf{R}_m|}, \quad (1)$$

Instead of this symmetric measure we want to find out if the reference trajectory is completely covered by another trajectory. Thus, we consider the following asymmetric similarity metric:

$$Sim(\mathbf{R}_a, \mathbf{R}_m) = \frac{|\mathbf{R}_a \cap \mathbf{R}_m|}{|\mathbf{R}_a|}, \quad (2)$$

$$Sim(\mathbf{R}_m, \mathbf{R}_a) = \frac{|\mathbf{R}_a \cap \mathbf{R}_m|}{|\mathbf{R}_m|}. \quad (3)$$

[4]earth-info.nga.mil/GandG/coordsys/grids/universal_grid_system.html

The first one shows what percentage of $\mathbf{R}_a$ is shared with $\mathbf{R}_m$ and the second shows what percentage of $\mathbf{R}_m$ is shared by $\mathbf{R}_a$. The way that we perform intersection operator is described in the following sections after we quantize the trajectories into cells.

3.1 Cell Approximation

In a preprocessing step, we generate a cell representation for a trajectory after it has been recorded. The Military Grid Reference Systems (MGRS) is an alpha-numeric system for expressing UTM/UPS coordinates. MGRS is used by NATO to locate points on earth. A single alpha-numeric value references a position that is unique for the entire earth (see Figure 2). MGRS is a projected coordinate system which uses a 2-dimensional Cartesian horizontal position orientation, so that locations are identified independently of vertical position. MGRS shares several characteristics with UTM such as the division of earth into projection zones and using easting and northing in meters within a designated zone. The main differences are that a MGRS zone is a 100km square within a UTM zone, whilst a UTM zone is usually 6 degrees in east-west and 8 degrees in north-south area and also that the notation of the areas is different. Based on the coordinate resolution, MGRS can define a grid with square cells with the length starting from 100km up to 10m or even 1m.

We approximate a trajectory $\mathbf{R} = \{(x_k, y_k)\}_{k=1}^{K}$ by a sparse binary matrix representation $\mathbf{C}$ where,

$$(\mathbf{C})_{ij} = \begin{cases} 1 & 0 < x_k - iL < L, 0 < y_k - jL < L \\ 0 & \text{Otherwise} \end{cases}, \quad (4)$$

where L stands for the cell length (25 meters in this paper) and indexes i and j span over in horizontal and vertical cells that trajectory R is residing inside. Figure 3 shows how the reference trajectory is approxi-

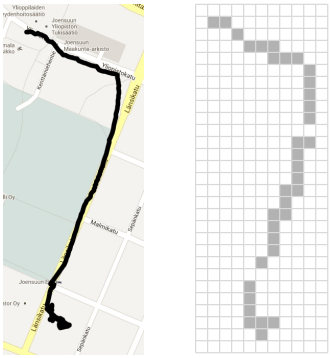


Figure 3: Example of a trajectory of 420 points being represented by 35 cells using the approximation in Equation 4. The cell representation is not continuous. The gaps appear because of the fixed cell size, variations in movement speed (or different sampling frequencies) and missing GPS locations. It is likely for such gaps to appear especially when users are moving by car, train or plane.

mated by cells. Generating the cell representation for a trajectory of average length of N points is done in $O(N)$ time.

3.2 Measuring Similarity

The similarity between two trajectories $\mathbf{R}_a, \mathbf{R}_m$ can now be calculated as:

$$Sim(\mathbf{R}_a, \mathbf{R}_m) = \frac{\|\mathbf{C}_a \odot \mathbf{C}_m\|_0}{\|\mathbf{C}_a\|_0}, \tag{5}$$

where $\mathbf{C}_a$ and $\mathbf{C}_m$ are the cell representations of $\mathbf{R}_a$ and $\mathbf{R}_m$, respectively, $\mathbf{C}_a \odot \mathbf{C}_m$ is a *Hadamard product* of two matrices $\mathbf{C}_a$ and $\mathbf{C}_m$ defined as $(\mathbf{C}_a \odot \mathbf{C}_m)_{ij} = (\mathbf{C}_a)_{ij} \cdot (\mathbf{C}_m)_{ij}$ and $\|\mathbf{C}\|_0$ represents the ℓ_0-quasinorm. In implementation, $\mathbf{C}_a$ and $\mathbf{C}_m$ are multiplied element by element and then we measure the number of non-zero elements. Figure 4 shows two sample trajectories being matched.

Assuming we have the cell representation $\mathbf{C}$ of a reference trajectory $\mathbf{R}$ we calculate the similarity for all trajectories in database in two steps. First, we find all the trajectories which share at least one cell with the reference trajectory. This has a time complexity of $O(N' \cdot (q + M'))$ where q represents the steps needed by the database system to perform the search ($N' \ll N$ and $M' \ll M$). In contrast to the average length N of a trajectory $\mathbf{R}$, we define $N' = \|\mathbf{C}\|_0$ as the number of non-zero elements in cell-approximated version of $\mathbf{R}$. In a similar way, M' indicates the number of other trajectories that share at least one cell with trajectory $\mathbf{R}$. Secondly, we calculate the trajectory similarity according to Equation (5) with a time complexity of $O(M' \cdot N')$. The overall complexity of the similarity

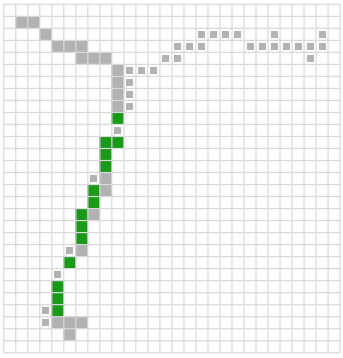


Figure 4: Matching two trajectories using the cell representation. The green cells denote the reference trajectory and the gray cells represent the other trajectory. The 'x' symbol is used to mark the cells shared by two trajectories; $Sim(\mathbf{C}_a, \mathbf{C}_m) = 40\%$ and $Sim(\mathbf{C}_m, \mathbf{C}_a) = 31\%$

scoring is $O(M' \cdot N')$ (assuming q constant by adding a proper indexing structure in the database).

In Figure 4 the straighforward application of the similarity scores yield similarity scores of 40% and 31% even though the trajectories seem to have more than 50% similarity by visual inspection. In the next subsections we analyze why this happens.

3.3 Interpolation

When the user is traveling fast or when recording frequency is low we notice gaps in the trajectory representation by cells. Gaps can also appear due to lack of GPS signal. Figure 5 shows three examples when different sized gaps appear in the cell representation of a trajectory. In cell approximation stage in Equation 4, we process the trajectory data points in the sequence they are recorded. In this way, the sequence of cells being detected as "1s" are used to determine if the next cell is connected to the current cell and find a potential gap in cell-approximation.

In order to fill the gap, the line equation between two cells is obtained from the start and end points as

$$j = f(i) = \frac{j_2 - j_1}{i_2 - i_1}(i - i_1) + j_1 \tag{6}$$

where i_1 and j_1 are the coordinates of one cell and i_2, j_2 are the coordinates of the other cell. The line in Equation 6 is then sampled by the cells that it is passing through and then set respective cell values as $(\mathbf{C}_{ij}) = 1$.

By performing interpolation, the trajectory similarity presented in Figure 4 is now updated as plotted in Figure 6. The similarity values are still below the visual expectations. The reason is that two cell representations may not overlap even though the trajectories are close to each other.

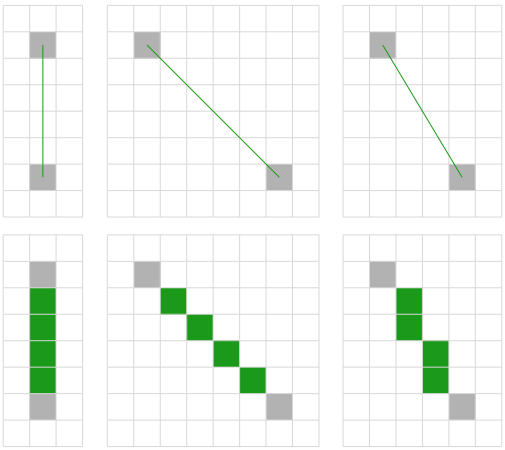


Figure 5: Interpolation between two cells in order to fill a gap; three example situations are depicted.

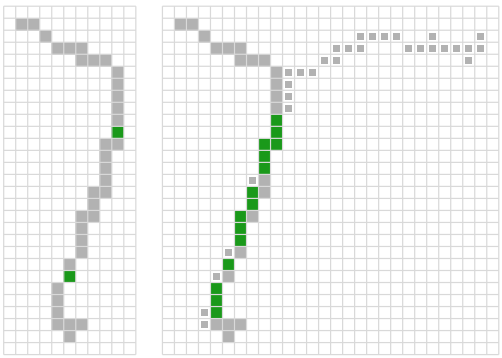


Figure 6: The trajectory having gaps is interpolated and the matching of the two trajectories becomes: $Sim(\mathbf{C}_a, \mathbf{C}_m) = 41\%$ and $Sim(\mathbf{C}_m, \mathbf{C}_a) = 33\%$.

3.4 Dilation

A frequent situation is that two nearby trajectory segments are evolving along each other in cell representation instead of overlapping. An example is provided in Figure 7 We solve this issue by applying *morphological dilation* on the trajectories and taking into account the neighbouring cells of a trajectory. We define $\mathbf{C}^d$ as a result of *binary dilation* of sparse binary representation $\mathbf{C}$ by *binary structure* $\mathbf{S}$ with

$$\mathbf{C}^d = \mathbf{C} \oplus \mathbf{S} = T(\mathbf{C} * \mathbf{S}), \tag{7}$$

where $\oplus$ defines the binary dilation and $*$ indicates the convolution operator. In the Equation 7, $T(\cdot)$ stands for *binarization transform* as

$$T((\mathbf{C} * \mathbf{S})_{ij}) = \begin{cases} 0 & 0 \leq (\mathbf{C} * \mathbf{S})_{ij} < 1 \\ 1 & \text{Otherwise} \end{cases} \tag{8}$$

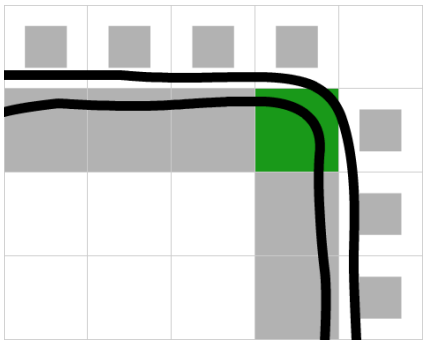


Figure 7: We see that two trajectories which are close enough to be considered similar can be represented by different cells. Only a single cell is shared by the cell representation of the two trajectories.

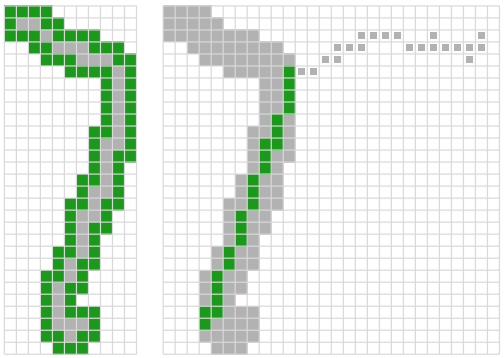


Figure 8: The reference trajectory is dilated and the matching of the two trajectories becomes: $Sim(\mathbf{C}_a, \mathbf{C}_m) = 64\%$ and $Sim(\mathbf{C}_m, \mathbf{C}_a) = 53\%$.

Figure 8 shows how a trajectory is dilated with the following structure

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{9}$$

Then the two trajectories are matched when one of the trajectories is dilated. The similarity score is now calculated with $\mathbf{C}_a$ and $\mathbf{C}_m^d$ as in Equation 5. Typically the number of cells used in the trajectory representations increases by a factor of 3 when dilation is applied.

4 RESULTS

We implement our method in a real-world application, as a prototype using the Mopsi project route analysis module [5]. We investigate issues that may appear

[5] cs.uef.fi/mopsi/?tab=routes&userId=13&routeId=137882401 9381&similarity=true

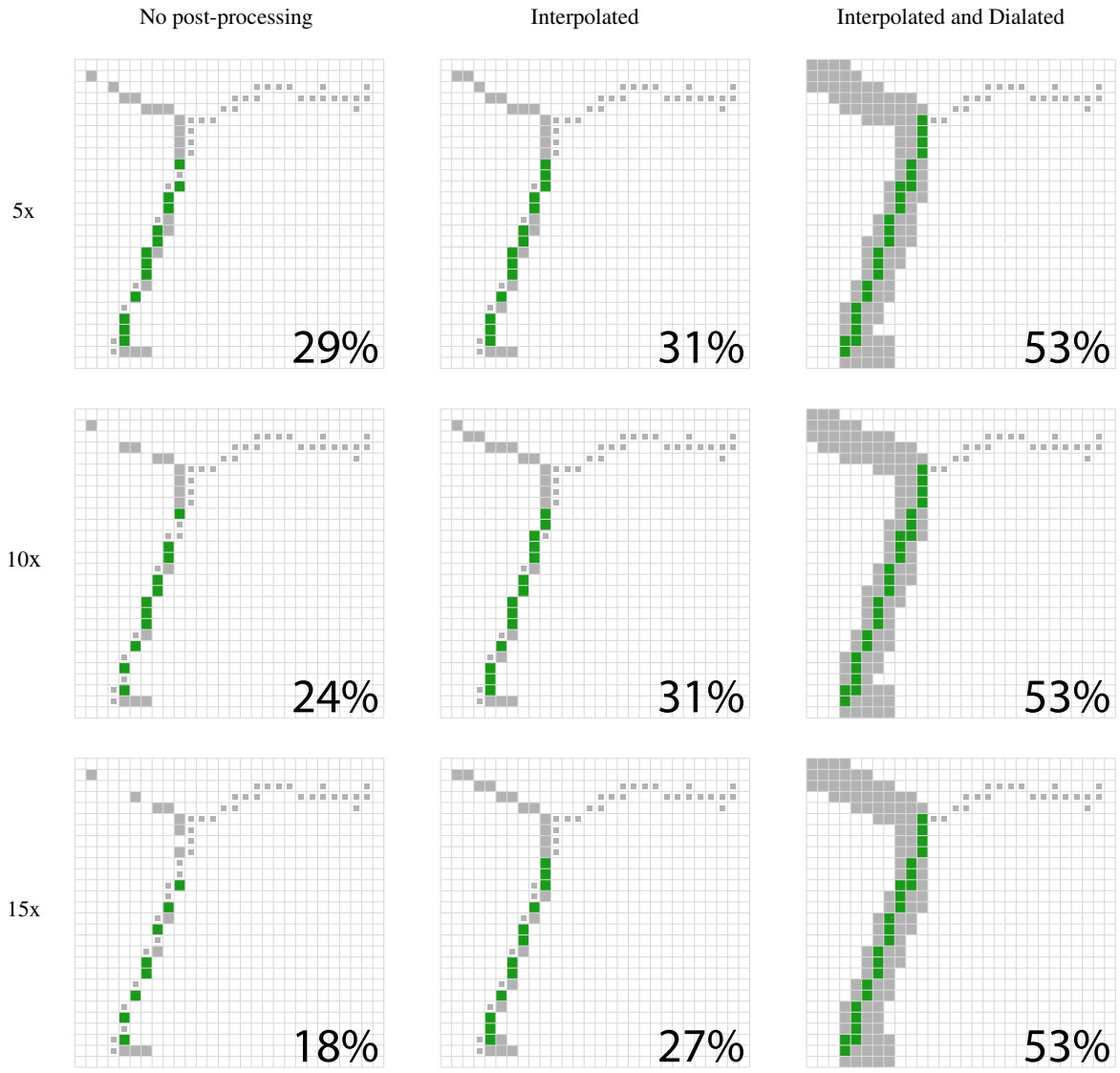


Figure 9: Simulating different sampling frequencies by subsampling the reference trajectory with a factor of 5x, 10x and 15x.

when collecting GPS trajectories in a practical application such as different sample rates, interpolation of collected points or breaks in the GPS signal caused by technical or environmental problems.

Firstly, as shown in Figure 9, we investigate how a different sampling frequency impacts the similarity score calculation. The reference trajectory is subsampled with factor f by only keeping every f^{th} element from the original trajectory. We notice that the interpolation step doesn't increase the similarity scores significantly. However, when followed by dilation, the similarity score indicates robustness against variations in sampling frequency which is a desired property for a trajectory matching procedure.

The other common issue while recording a trajectory is loss of location information for a brief period of time. This can happen, for example, if the user goes through a building, a tunnel or simply due to device software error. We simulate this behavior and see how the similarity scoring is affected in Figure 10. When removing 90 points we notice that the similarity score has dropped even when using interpolation and dilation. This happened because we removed a significant amount of subsequent points (20% of the trajectory). Interpolation does not have enough information to reconstruct the trajectory appropriately and consequently, loss of many data points in a trajectory is detrimental for similarity calculations.

The proposed method is implemented in two steps for real-world application: the preprocessing step,

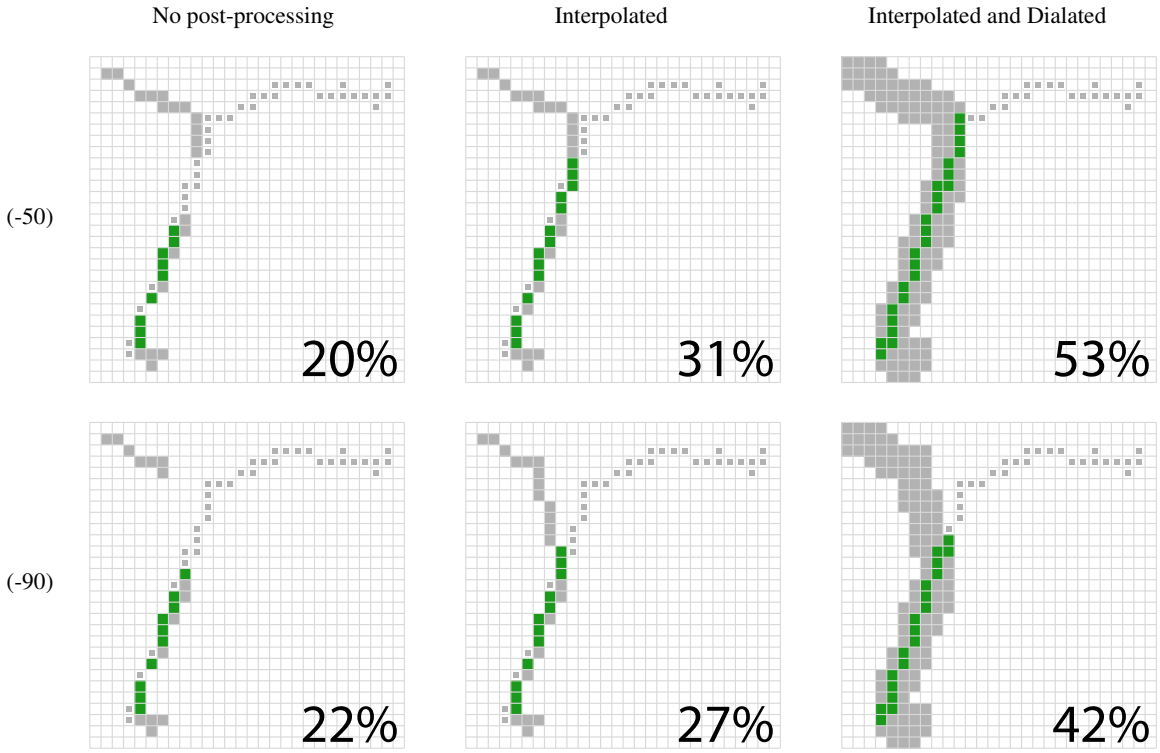


Figure 10: Simulating loss of GPS signal by removing 50 and 90 sequential points from the reference trajectory.

done when a new trajectory is added into the system and the similarity score calculation step, performed when searching all the similar trajectories of a given trajectory. When not using interpolation or dilation the time complexity for the preprocessing step is $O(M \cdot N)$ for M trajectories of average length N points. The similarity score calculation has a time complexity of $O(M' \cdot N')$. After interpolation is applied there will be an increase on the N' and M' parameters which increase, however, stay at the same order of magnitude. N' increases by the number of cells added trough interpolation and M' increases by the number of trajectories that share at least one cell with interpolated trajectory. The dilation stage increases the N' and M' parameters once more. N' typically increases by a factor of 3 and M' grows by the number of trajectories that share the cells that are added to the representation as a result of dilation. The overall complexity for M trajectories in the database is governed by $O(M \cdot N)$ for cell approximation and $O(\alpha \cdot M \cdot M' \cdot N')$ for similarity score calculation including interpolation and dilation ($\alpha \approx 6$, $M' \ll M$, $N' \ll N$). The similarity cell approximation complexity of $O(M \cdot N)$ is negligible compared to $O(\alpha \cdot M \cdot M' \cdot N')$ for score calculation. Hence, the overall computational complexity of the proposed approach is dominated by $O(\alpha \cdot M \cdot M' \cdot N')$ which is comparably much less than $O(M^2 \cdot N^2)$ for other similarity metrics presented in section 1.

5 CONCLUSIONS

We presented a method for computing similarity between trajectories in a large data collection. Because trajectories are likely to have different speed profile and missing points, interpolation and dilation techniques are employed before the scoring. We have demonstrated that the method is robust except when many points are removed and dramatically affect the structure of a trajectory. In that situation there is simply not enough information to rebuild the path and provide correct similarity values. The method was implemented in Mopsi, where for a given trajectory we display a list of similar paths in reverse order of the similarity scores.

REFERENCES

Agrawal, R., Faloutsos, C., and Swami, A. (1993). *Efficient similarity search in sequence databases*. Springer.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time

warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

Chan, K.-P. and Fu, A. W.-C. (1999). Efficient time series matching by wavelets. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 126–133. IEEE.

Chen, L., Özsu, M. T., and Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502. ACM.

Chen, M., Xu, M., and Fränti, P. (2012a). Compression of gps trajectories. In *Data Compression Conference (DCC), 2012*, pages 62–71. IEEE.

Chen, M., Xu, M., and Fränti, P. (2012b). A fast $O(N)$ multiresolution polygonal approximation algorithm for GPS trajectory simplification. *IEEE Transactions on Image Processing*, pages 2770–2785.

Fränti, P., Chen, J., and Tabarcea, A. (2011). Four aspects of relevance in location-based media: content, time, location and network. In *Web Information Systems and Technologies (WEBIST'11), International Conference on*, pages 413–417.

Fränti, P., Tabarcea, A., Kuittinen, J., and Hautamäki, V. (2010). Location-based search engine for multimedia phones. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 558–563. IEEE.

Frentzos, E., Gratsias, K., Pelekis, N., and Theodoridis, Y. (2007a). Algorithms for nearest neighbor search on moving object trajectories. *Geoinformatica*, 11(2):159–193.

Frentzos, E., Gratsias, K., and Theodoridis, Y. (2007b). Index-based most similar trajectory search. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 816–825. IEEE.

Güting, R. H., Behr, T., and Xu, J. (2010). Efficient k-nearest neighbor search on moving object trajectories. *The VLDB Journal*, 19(5):687–714.

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Cambridge Univ Press.

Hu, N. and Steenkiste, P. (2006). Quantifying internet end-to-end route similarity. In *Passive and Active Measurement Conference*, volume 2006, pages 101–110.

Lange, D. and Naumann, F. (2011). Efficient similarity search: arbitrary similarity measures, arbitrary composition. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1679–1688. ACM.

Mariescu-Istodor, R. (2013). Detecting user actions in MOPSI. Master's thesis, University of Eastern Finland.

Ni, J. and Ravishankar, C. V. (2007). Indexing spatio-temporal trajectories with efficient polynomial approximations. *Knowledge and Data Engineering, IEEE Transactions on*, 19(5):663–678.

Pelekis, N., Kopanakis, I., Kotsifakos, E. E., Frentzos, E., and Theodoridis, Y. (2011). Clustering uncertain trajectories. *Knowledge and Information Systems*, 28(1):117–147.

Tabarcea, A., Wan, Z., Waga, K., and Fränti, P. (2013). O-mopsi: Mobile orienteering game using geotagged photos. pages 300–303.

Vlachos, M., Gunopulos, D., and Kollios, G. (2002a). Robust similarity measures for mobile object trajectories. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 721–726. IEEE.

Vlachos, M., Kollios, G., and Gunopulos, D. (2002b). Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 673–684. IEEE.

Waga, K., Tabarcea, A., Chen, M., and Fränti, P. (2012). Detecting movement type by route segmentation and classification. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, pages 508–513. IEEE.

Waga, K., Tabarcea, A., and Fränti, P. (2011). Context aware recommendation of location-based data. In *System Theory, Control, and Computing (ICSTCC), 2011 15th International Conference on*, pages 1–6. IEEE.

Waga, K., Tabarcea, A., Mariescu-Istodor, R., and Fränti, P. (2013). Real time access to multiple GPS tracks. pages 293–299.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fMRI time-series revisitedagain. *Neuroimage*, 2(3):173–181.

Yanagisawa, Y., Akahani, J.-i., and Satoh, T. (2003). Shape-based similarity query for trajectory of mobile objects. In *Mobile data management*, pages 63–77. Springer.

SocialSearch
A Social Platform for Web 2.0 Search

Claudio Biancalana, Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti
Department of Engineering, Artificial Intelligence Laboratory, Roma Tre University,
Via della Vasca Navale, 79, 00146 Rome, Italy
{*claudio.biancalana, gaspare, micarel, gsansone*}*@dia.uniroma3.it*

Keywords: Query Expansion, Social Bookmarking Services, Personalization.

Abstract: In the last decade, social bookmarking services have gained popularity as a way of annotating and categorizing a variety of different web resources. The idea behind this work is to exploit such services for enhancing traditional query expansion techniques. Specifically, the system we propose relies on three-dimensional co-occurrence matrices, where the further dimension is introduced to represent categories of terms sharing the same semantic property. Such categories, named *semantic classes*, are related to the folksonomy mined from social bookmarking services such as *Delicious*, *Digg*, and *StumbleUpon*. The paper illustrates a comparative experimental evaluation on real datasets, such as the one collected by the Open Directory Project and the TREC 2004. We also include the results of a specific disambiguation analysis aimed to evaluate the effectiveness of our approach in comparison with state-of-the-art techniques when satisfying queries characterized by polysemic and ambiguous terms.

1 INTRODUCTION

The Social Semantic Web combines together the core principles of the Semantic and Social Web: it includes, on the one hand, the idea of associating a semantic description with web resources for enabling machines to access and process them, on the other hand, the idea of exploiting social content information for that purpose. This development, however, leads to the need to revise the classical techniques for the traditional Web (Micarelli et al., 2006; Lops et al., 2007; Gentili et al., 2001; Gasparetti and Micarelli, 2003), as they could not be more efficient in the new Web design.

Automatic query expansion (QE) is a well-known technique for enabling users to better characterize their search domain by supplementing the original query with additional terms that are somehow linked to the frequency of the term the user specified in his query (Bai et al., 2005). This method can significantly improve the performance of information retrieval systems. However, traditional QE techniques, even those providing users with personalized results, may suffer from some drawbacks if extended to the Social Semantic Web. In particular, additional terms can (i) be simple synonyms, (ii) not consider the existence of different lexicons, given that each user has his own custom dictionary. As a result, QE process may fail to contextualize the research domain of interest if multiple users annotate the web content.

Our research objectives include (i) to find a solution to the lack of expression of the candidate terms for query expansion, (ii) to customize the search results taking into account the semantic domain of the user interests, (iii) more generally, to explore novel approaches combining semantic, social, and adaptive aspects.

The proposed system - named *SocialSearch* - is an extension of the traditional QE techniques, which are based on the computation of two-dimensional co-occurrence matrices (Biancalana and Micarelli, 2009; Biancalana et al., 2012). Our approach makes use of three-dimensional co-occurrence matrices, where the added dimension is represented by *semantic classes* (i.e., categories comprising all the terms that share a semantic property) related to the folksonomy extracted from social bookmarking services (Musto et al., 2009) such as *Delicious* [1], *Digg* [2] and *StumbleUpon* [3]. These web sites allow users to store, organize, share, and search bookmarks associated with web resources, through the input of additional data

[1] delicious.com
[2] digg.com
[3] www.stumbleupon.com

(e.g., tags or short summaries), freely available to the entire community of users. In our approach, the expansion process takes place by analyzing multiple occurrences divided into categories related to semantic classes, which are analyzed in the folksonomy. The full process is entirely transparent to the user, implicitly occurring based on his past choices related to the terms of the submitted queries and the corresponding visited pages. The input queries are analyzed according to collected data, and if they reflect the interests already shown by the user in previous searches, the system returns different QEs, before performing the search phase. All of these QEs are related to the terms of the user query, but each of them involves a different semantic field. The final results are displayed in different blocks - each one classified through keywords - thereby supporting the user in determining what is most relevant to him (Acampora et al., 2010). This way, our system provides the contextualization and categorization of the information by analyzing and extracting the semantic domain of the user interests.

A comparative analysis of our findings with those obtained through some state-of-the-art techniques, such as relevance feedback, shows that our approach is able to achieve better results. This reveals that our system can offer a stronger correlation with the actual user interests, which confirms the validity and usefulness of their categorization in semantic classes.

The rest of the paper is structured as follows. Section 2 reviews some related works, Section 3 illustrates the system architecture. The main algorithms are detailed in Section 4, while Section 5 is devoted to the presentation and discussion of the experiments we performed. Finally, Section 6 concludes the paper and highlights some future directions.

2 RELATED WORK

Automatic query expansion (QE) has been widely used in Information Retrieval (Carpineto and Romano, 2012; Biancalana et al., 2013b). Among the various QE approaches proposed in literature, some of them take advantage of the implicit relevance feedback through pseudo-relevance feedback (PRF) (Manning et al., 2008). All these methods follow the basic assumption: documents classified higher by an initial search contain many useful terms that can help discriminate relevant documents from irrelevant ones. Despite the large number of studies, a crucial issue is that the expansion terms identified through traditional methodologies from the pseudo-relevant documents may not be all useful (Cao et al., 2008).

Bilotti et al. (Bilotti et al., 2004) analyze the effect of some QE approaches on document retrieval in the context of question answering, mainly targeted to the so-called "factoid" questions, namely, fact-based, natural language questions that usually can be answered by a short noun phrase. More specifically, the authors describe a quantitative comparative analysis between two different strategies for tackling term variation: i) employing a stemming algorithm at indexing time, or ii) carrying out a morphological query expansion at retrieval time. The findings show that, when compared to the baseline (no stemming nor expansion), stemming yields a lower recall, while morphological expansion results in higher recall. However, higher recall is paid at the cost of retrieving more irrelevant documents and ranking relevant documents at lower positions.

One of the failure reasons of the query expansion has been identified in the lack of relevant documents in the local collection. Consequently, some works advance the use of an external resource for query expansion in order to improve the effectiveness of query expansion, such as thesaurus (Nanba, 2007), Wikipedia (Xu et al., 2009), key-phrases from corpus of documents (Biancalana et al., 2013a; Biancalana et al., 2011), browsed web pages (Gasparetti et al., 2014) and search engine query logs (Cui et al., 2003). Abouenour et al. (Abouenour et al., 2010) point out that the adoption of a thesaurus, typically constructed through statistical techniques, poses several drawbacks. First of all, the construction of a thesaurus is time-consuming because of the great deal of data to process. Effective semantic QE techniques can also rely on ontologies instead of thesauri. Indeed, ontologies describe both semantic and concept relations, and enable semantic reasoning as well as cross-language information retrieval. The authors specifically deal with the enhancement of question answering in Arabic, a complex language for its peculiarities. They propose an approach that implements a semantic QE based on the WordNet [4]ontology in Arabic. As a result, the described QE method bears the following semantic relations: synonymy, hypernymy (supertypes), hyponymy (subtypes), and the Super Upper Merged Ontology (SUMO) [5] concept definition. SUMO is a top-level ontology that defines general terms and can be used as a foundation for middle-level and more specific domain ontologies. The documents retrieved through the previous process are then re-ranked using a structure-based approach based on the Distance Density n-gram model.

Recently, several authors have focused on social

[4]wordnet.princeton.edu

[5]www.ontologyportal.org

annotations as external resource, largely motivated by their increasing availability through many Web-based applications. Among these, Carman et al. (Carman et al., 2009) explore how useful tag data may be to improve search results, but they focus primarily on data analysis rather than retrieval experiments. Zhou et al. (Zhou et al., 2012) propose a query expansion framework relied on user profiles extracted from the annotations and resources bookmarked by users. The main difference with our approach is that the selection of expansion terms for a given query is not based on semantic classes, but on the assumption that they are likely to have similar weightings influenced by the documents best ranked for the original query.

3 SYSTEM ARCHITECTURE

In this section we present the architecture of the system we propose (see Fig. 1), describing the functionalities of each module and the modalities which they actively collaborate through.

- **Interface:** the mail role of the this module is readdressing external requests to the specialized modules and processing the achieved results so as to show them in a more understandable form;
- **Expansion:** after the user has submitted his search query, this module is responsible of the query expansion process. To perform multiple expansions, this module has to access the user interests stored in the user model;
- **Search:** this module is in charge of the real search process, receiving (possibly expanded) queries in input and returning the corresponding results;
- **Persistence:** all the necessary information is retained in this module: login data, encountered terms (both before and after stemming), tags, co-occurrence values between terms, tag relevance, and URLs of documents visited by the user; it interacts mainly with the interface (for user login and saving URLs) and the user model (for data needed for the construction and analysis of the user model);
- **UserModel:** this is the largest module in that it has to constantly update the user profile realized as a three-dimensional co-occurrence matrix. The interaction with the persistence module is the first step for achieving data (visited URLs and corresponding queries) from which to infer information for the model update. Before the necessary processing, this module makes use of two other sub-modules: Parser and TagFinder;
  - **Parser:** its main role is to filter out the unnecessary information related to the user interests collected by the system, and to provide the user model with a sorted set of terms for the three-dimensional matrix computation. It includes parsing functionalities (i.e., filtering the HTML pages visited by the user), stemming, and stop-word removal;
  - **TagFinder:** it is devoted to the search of tags to be associated with the pages visited by the user. It interacts with external resources (social bookmarking services) to extract complete tags of a relevance index, in order to supply them to the user model.

Results obtained in each search session are then presented to the user so as to underline the different semantic categories of each group of them. The search of the tags associated with the pages visited by the user is performed by analyzing the information provided by main sites that offer social bookmarking services. In this case, data collection occurs directly by parsing the HTML pages containing the necessary information. In order to model the user visits, the system employes matrices based on co-occurrence at the page level: terms highly co-occurring with the issued keywords have been proven to increase precision when appended to the query (Biancalana et al., 2009). The generic term t_x is in relation with all other n terms t_i (with $i = 1,\ldots,n$) according to a coefficient c_{xi} representing the co-occurrence measure between the two terms. In a classical way, we can construct the co-occurrence matrix through the Hyperspace Analogue to Language approach (Burgess and Lund, 1995): once a term is given, its co-occurrence is computed with n terms to its right (or its left); in particular, given a term t and considered the window f_t of n terms w_i to its right $f_t = \{w_1,\ldots,w_n\}$, we have $cooc(t,w_i) = \frac{w_i}{i}$, $i = 1\ldots,n$. A pair (a,b) is equal to pair (b,a), that is, the co-occurrence matrix is symmetrical. For each training document a co-occurrence matrix is generated, whose lines are then normalized to the maximum value. The matrices of the single document are then summed up, thus generating one single co-occurrence matrix representing the entire corpus.

The limit of this structure lies in the latent ambiguity of collected information: in presence of polysemy of the terms adopted by the user, the result of the query expansion risks to misunderstand the interests, so leading to erroneous results. In order to overcome this problem, in our system the classical model of co-occurrence matrix has been extended. The user model consists of a three-dimensional co-occurrence matrix. Each term of the matrix is linked to an intermediate level containing the relative belonging classes, each

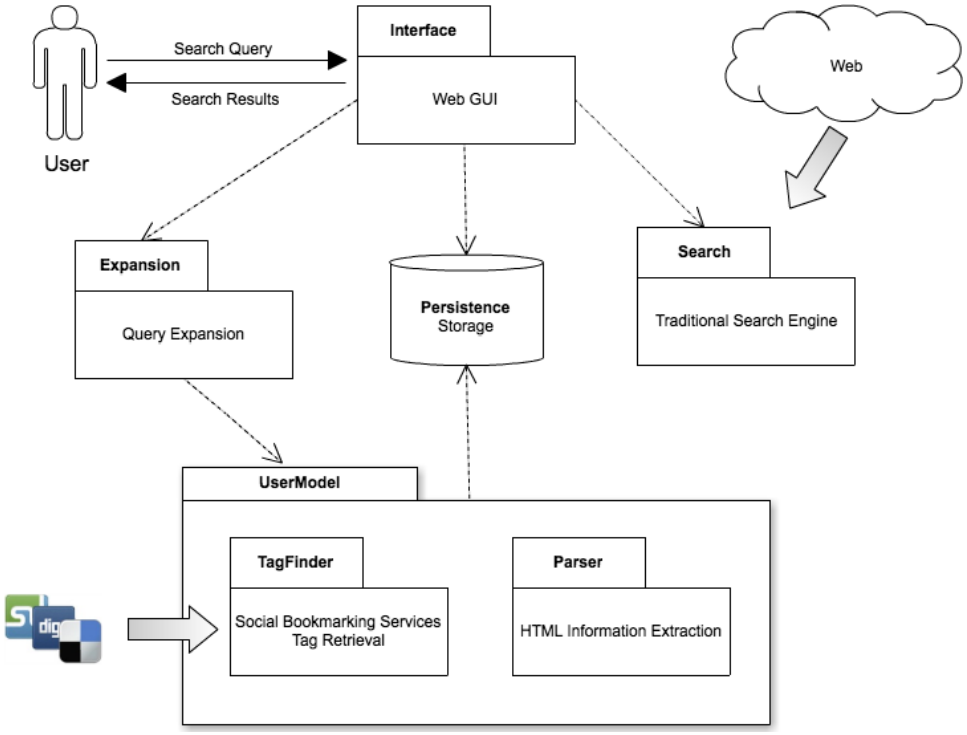


Figure 1: The system architecture.

accompanied by a relevance index. This way, each term is *contextualized* before being linked to all the other terms present in the matrix, and led to well determined semantic categories that are identified by tags.

4 SOCIAL SEARCH

In this section we describe in detail the two main algorithms of our approach. The former is designed for the user model creation and update (discussed in Section 4.1), the latter for the query expansion process (discussed in Section 4.2). With reference to the pseudocode shown below, we notice that the co-occurrence matrix is represented through a map of maps for encoding knowledge and connecting such knowledge to relevant information resources. Maps of maps are organized around *topics*, which represent subjects of interest; *associations*, which express relationships between the subjects; and *occurrences*, which connect the subjects to pertinent information resources.

4.1 User Model Creation and Update

The creation and update of the user model are based on the pages chosen by the user while searching. Starting with an empty model, every time the user clicks on a result after typing a search query, the system records the visited URL, together with the query originally submitted for the search. Our system performs the analysis of the visited URLs in incremental way, according to the following algorithm (see Algorithm 1, where capital deltas (Δ) denote comments):

- a temporary map M is initialized, where it is possible to store the extracted data, before updating the pre-existent model (empty at first execution). The map keys are the encountered tags, the values are the relative two-dimensional co-occurrence matrices;
- for each visited URL, the corresponding HTML page is obtained, from which the textual information is extracted through a parser, as a list of terms;
- the list of terms is filtered in order to eliminate stopwords (i.e., all those terms that are very frequent in all documents, so irrelevant to the creation of the user model);
- the list of terms undergoes a stemming process by means of the Porter's algorithm (Porter, 1980). At the same time the system retains the relations between stemmed terms and original terms;
- the co-occurrence matrix corresponding to the most relevant k_{term} keywords is evaluated. The relevance is measured by counting the occur-

Algorithm 1: User Model Creation and Update.

```
begin
    Δ Initialize the global co-occurrence matrix M (map of maps);
    M ← Map([]);
    Δ Analyze training documents;
    for (doc, query) in D do
        Δ Parse the document (stemming and stopword removal);
        doc = parse(doc);
        Δ Initialize the co-occurrence matrix of different terms;
        terms ← Map([]);
        Δ Compute the co-occurrence value of every term;
        terms = frequency_occurrences(doc);
        Δ Initialize the co-occurrence matrix of document;
        co_occ ← Map([]);
        Δ Compute the co-occurrence matrix of document;
        co_occ = co_occurrences(terms);
        Δ Get the site list of social bookmarking for tag search;
        sites = get_social_bookmarking_sites();
        Δ Initialize URL list tags;
        tags ← Set([]);
        Δ Retrieve tags by URL;
        for i = 0; i < sites.size() & tags.size() = 0; i++ do
            tags = retrieve_tags(url, sites[i]);
        Δ Update the matrix M;
        update(M, tags, terms);
    Δ Initialize all terms in documents;
    all_terms ← Set([]);
    Δ Get unique terms set;
    all_terms = get_term_set(M);
    Δ Get subset of user model;
    user_matrix ← get_user_matrix(all_terms);
    Δ Update user model by the intermediate matrix;
    update(user_matrix, M, all_terms);
    Δ Store updated user model;
    save(user_matrix);
```

rences within the document itself, with the exception of terms used in the query (retained by the system together with the corresponding URL), to which is assigned the maximum weight;

- tags concerning the visited URLs are obtained by accessing different sites of social bookmarking. Each extracted tag has a weight which depends on its relevance (i.e., the number of users which agree to associate that tag to the visited URL);
- the update of the temporary map M is performed by exploiting all the information derived from the co-occurrence matrix and the extracted tags in a combined fashion. For each tag_i the system updates the co-occurrence values just calculated, according to the tag relevance weight. After that, the vectors M_{tag_i,t_i} related to each term t_i are updated by inserting the new (or summing to the previous) values;
- the set $terms$ is calculated, which contains all the terms encountered during the update of the temporary map M;
- from the persistence module a subset UM_{terms} of the user model is obtained as a three-dimensional matrix of co-occurrences, corresponding only to the terms contained in $terms$;
- the matrix UM_{terms} is updated with the values of M. For each t_i belonging to $terms$, the set of keys ($tags$) is extracted from M, which points to values corresponding to t_i. For each tag_i belonging to $tags$, the vector M_{tag_i,t_i} is added to the pre-existent vector UM_{t_i,tag_i}, updating the values for the terms already present and inserting new values for the terms never encountered.

4.2 Query Expansion

The query expansion process is performed beginning from the original terms entered into the search engine by accessing the information collected in the user model. The result is a set of expanded queries, each of them associated with one or more tags. This way, it is possible to present the user with different subgroups of results grouped in categories. Using low level boolean logic, every expansion assumes the following form:

$$(t_{11}\ OR\ \ldots\ OR\ t_{1x})\ AND\ (t_{21}\ OR\ \ldots\ OR\ t_{2x})\ \ldots\ AND\ (t_{y1}\ OR\ \ldots\ OR\ t_{yx})$$

where t_{yx} represents the generic term x corresponding to the stemmed root y. The different terms coming from the same root undergo *OR* operation amongst them, since the result has to contain at least one of them. (see examples in Table 1).

Table 1: Example of multiple expansions.

Original query	Categorization tags	Expansions
amazon	**e-commerce, shopping:**	buy **AND** (books **OR** book) **AND** amazon
amazon	**nature:**	(rivers **OR** river) **AND** amazon

The algorithm of multiple expansion is the following (see Algorithm 2):

- let us suppose that the query Q is given, which consists of n terms q_i (with $i = 1,\ldots,n$). For each of them the system evaluates the corresponding stemmed term $q_i^{'}$, so obtaining the new query $Q^{'}$ as a new result;

- for each term belonging to $Q^{'}$, the corresponding two-dimensional vector q_i is extracted from the three-dimensional co-occurrence matrix. Each of those vectors may be viewed as a map, whose keys are the tags associated with the terms $q_i^{'}$ (which have a relevance factor), and the values are themselves *co-occurrence vectors* between $q_i^{'}$ and all the other encountered terms;
- for each encountered tag the relevance factor is recalculated, adding up the single values of each occurrence of the same tag in all two-dimensional vectors. This way, the result is a vector T in which tags are sorted according to the new relevance factor;
- amongst all tags contained in T, only the higher k_{tag} are selected and considered for the multiple expansions;
- for each selected tag t_i the vector sum_{t_i} is computed, which represents the sum of the co-occurrence values of the three-dimensional matrix, corresponding to all terms $q_i^{'}$ of the query $Q^{'}$;
- for each vector sum_{t_i}, the most relevant terms k_{qe} (corresponding to higher values) are selected. Combining the extracted terms with those of the query Q, a new query $EQ^{'}$ (made up of stemmed terms) is initialized;
- for each expanded query $EQ^{'}$, the corresponding query EQ is calculated through the substitution of stemmed terms with all the possible original terms stored into the system, exploiting the boolean logic according to the scheme previously shown;
- the query EQ and the original tag t_i are entered into the map M_{EQ}, whose keys are expanded queries and values are sets of tags. If M_{EQ} already contains an expanded query identical to the input one, the tag t_i is added to the corresponding set of tags.

Algorithm 2: Multiple Query Expansion.

```
begin
    Δ Initialize the query to be expanded (a list of n terms);
    query ← [q1, q2, ..., qn];
    Δ Stemming of query terms;
    query ← stemming(query);
    Δ Get the subset of the user model related to the query;
    user_matrix = get_user_matrix(query);
    Δ Initialize the tag map for multiple query expansion;
    expansion_tags ← Map([]);
    Δ Compute tags for multiple expansion;
    expansion_tags = find_expansion_tags(query, user_matrix);
    Δ Initialize the expanded query map related to tags;
    exp_queries ← Map([]);
    Δ Compute expanded queries for every tag;
    for (tag, ranking) in expansion_tags do
        Δ Compute the expanded query by choosing most
        relevant terms;
        exp_query = select_relevant_terms(query, user_matrix);
        Δ Enter the result in the expanded query map;
        insert_expanded_query(exp_query, tag, ranking, exp_queries);

    return exp_queries;
```

5 EVALUATION

We now present the experimental results of the proposed approach. Specifically, we describe a comparative evaluation analysis between *SocialSearch*, our social-based search engine, and some state-of-the-art techniques.

A number of different aspects must be evaluated in order to assess the real effectiveness of search engines, such as index coverage, search capabilities, presentation, and user effort in seeking tasks. In this evaluation, we are particularly interested in the standard relevance measures to evaluate the efficacy of the retrieval of web documents and the quality of the results. Several relevant factors make this comparative analysis somewhat difficult. Personalized search aims at enhancing user interaction by understanding the user needs, the context, and the applications and information being used, typically across a wide set of user goals. Usage data that might be of potential interest for recognizing and assessing information consumption patterns of each user and the various information foraging strategies must be accurately collected. Moreover, personalization is influenced by the selection of particular topics on which the evaluation is to be performed. It can create an authoring bias where the topics selected by a group of peers influence the relative results of one approach when compared with others. For example, one approach might exploit a topic characterized by a wealth of documents and references, while a different one is critically affected by the presence of several polysemous words in the query set. In spite of these issues, implementing an experimental evaluation of personalized approaches in a real setting is still the most significant method to measure the scalability and the overall quality of search effectiveness, in terms of both coverage and accuracy of the produced search results. While coverage measures the ability of engines to produce all the references that are likely to be visited by the user, accuracy is essential in evaluating the quality of such references.

Five different search engines have been included in the comparative analysis: Google (denoted simply as *Google* in figures), the personalized version of Google (*PersGoogle*), a query expansion search engine based on co-occurrence data (*CoOcc*), a traditional search engine with Relevance Feedback (*RF*), and our system (*SocialSearch*). In the first personalized version of Google back in 2004, the search engine showed a directory like category drop-down menu, where users could select the categories that matched their interests. During the search process, the search engine adapts the results according to each user needs, assigning a higher score to the resources related to what the user has seen in the past. A slider in the graphic user interface allows the user to control the level of personalization in the results. For example, if the user earlier chose the category of *Computers* as one of his interests, results such as *Apple*, *Acer* or *HP* would rank among the first positions. Unfortunately, no details or evaluations are presently available for the algorithms exploited for the re-ranking process, except the ones contained in the patent application filed in 2004 (Zamir et al., 2004). Our comparative evaluation takes into account the current version of personalized Google. It basically reorders the search results based on gathered usage data, such as previous queries, web navigation behavior and, possibly, visited sites that serve Google ads, computers with Google Applications installed, such as Desktop Search and personal information, which may be implicitly or explicitly provided by the user.

Relevance feedback aims at modifying the initial query using words extracted from top-ranked or identified relevant documents. If both documents and queries are represented in a vector space model (Salton and Buckley, 1997), the Rocchio feedback approach alters the initial query by combining the vectors of the relevant documents increasing the recall of the search engine, and possibly its precision as well (Manning et al., 2008).

Query expansion based on co-occurrences is a well-known approach that collects the correlations between pairs of terms in a given corpus. It is a straightforward approach that limits the computational complexity through the idea of associating contexts to the current user needs. The two fundamental problems of information retrieval, namely, synonymy and polysemy, are addressed during the construction of the query vector. Ambiguous words have only one lemma for all their meanings. If one meaning is mentioned in a query, the documents in which the term appears with the other meanings are also retrieved and estimated as closer to the query. In case of polysemy there will be terms associated to more than one meaning, but if the query is composed by a number of keywords, the intended meaning is more likely to be referenced. These terms and their associated terms will form a cluster, which is associated to the intended meaning and outweighs the unintended meanings. Several studies in the literature have proven the effectiveness of this approach, but have also raised some doubts on its real improvements in the performance of document retrieval systems, because of the following potential issues:

- Weighting terms that occur more frequently in the whole dataset, so favoring the more popular (see, for example, (Peat and Willett, 1991));
- Expanding each single term in the query in isolation, ignoring the potential meaning of the all terms as a whole;
- Co-occurrences data extracted from small collections of documents;
- Collection of documents not including relevant concepts and information during the query expansion.

In order to play down those issues mainly related to the documents selected for the initial dataset, the co-occurrence matrix used for expansion is built on the corpus of documents retrieved during the learning process. In this way, it is certain that enough relevant documents for the expansion are included and there are less chances to see several common terms that cover several different topics of interests. The comparative analysis consists in the following two evaluations:

- TREC corpus-based evaluation;
- ODP corpus-based evaluation;

Corpus-based evaluations have the advantage of showing a zero test-retest variability if the same closed corpus is employed in future experiments that include different approaches. We also include a specific disambiguation analysis in order to measure the efficacy of the search engines to tackle queries characterized by polysemic and ambiguous terms.

5.1 TREC Corpus-based Evaluation

In the first evaluation, we consider the TREC [6] 2004 Robust Track on TREC disks 4 and 5. It contains over 500K documents, a subset of them marked relevant or irrelevant according to a given topic. On average, each document consists of 467 terms. All the 249 queries are included in the evaluations. The approaches considered in this evaluation are *RF*, *CoOcc*,

[6] trec.nist.gov/data.html

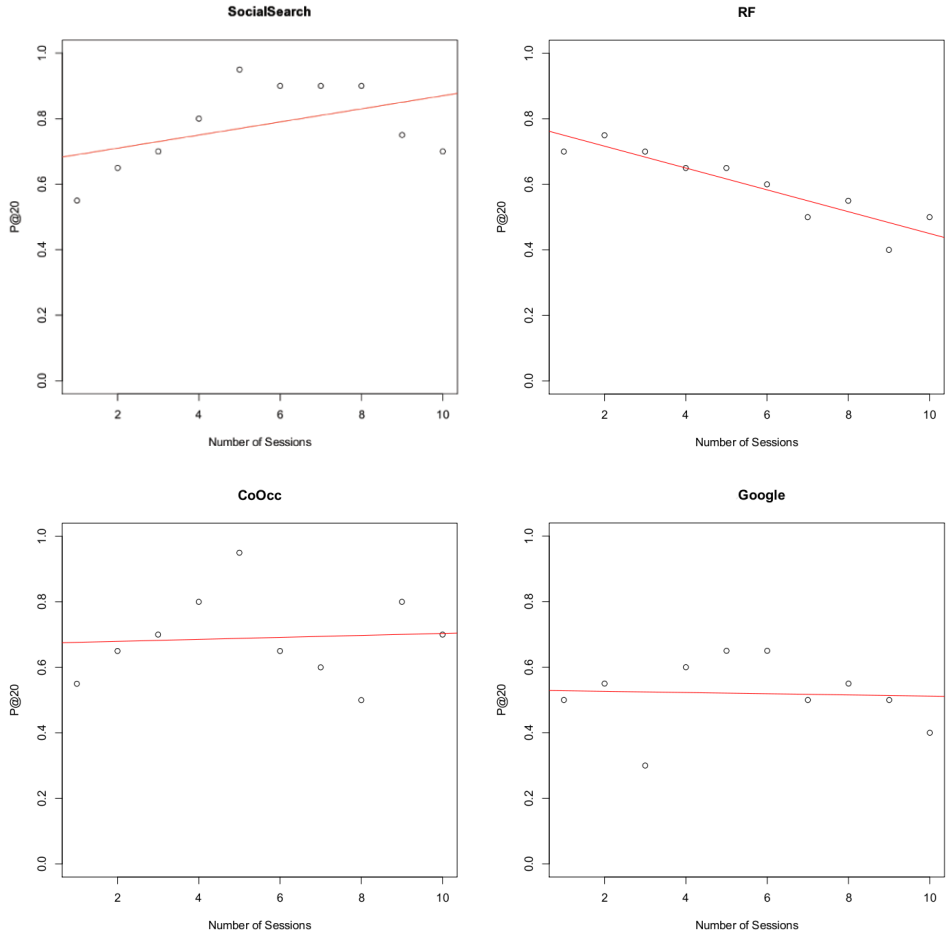


Figure 2: P@20 average values after a certain number of feedbacks.

Google, and *SocialSearch*. The closed nature of this corpus has not allowed us to include *PersGoogle* in this comparative analysis as well. The precision at 20 (P@20) measures the performance of the retrieval. It evaluates the fraction of the retrieved documents that are relevant to the user information needs. The average number of result pages viewed by a typical user for a query is 2.35 (Jansen et al., 2000), and a more recent study (Jansen et al., 2005) reports that about 85.92% of users view no more than two result pages. For these reasons, the precision is evaluated at a given cut-off rank, considering only the top 20 results returned by the system. Figure 2 shows the P@20 after collecting a certain number of feedbacks (Biancalana et al., 2013b).

Google approach shows the worst outcomes with a low average precision. This is an expected result because *Google* does not exploit the suggestions that feedbacks might provide. Better average outcomes are obtained by employing the relevance feedback, even though the slope of the linear model of data is negative. That is to say that the amount of information collected by means of the relevance feedback negatively affects the precision by including irrelevant keywords during the expansion of the queries. Better outcomes are obtained through both *CoOcc* and *SocialSearch* approaches. It must be noted how several web references included in the corpus do not find a correspondence in the sites of social bookmarking services such as *Delicious*. For this reason, *SocialSearch* is put in a unfavorable position in comparison with *CoOcc* trained on the collection of documents related to the relevant topics. The same issue also affects the ODP corpus-based evaluation (see Sect. 5.2).

5.2 ODP Corpus-based Evaluation

Our goal is to build profiles of users that show interests in some specific topics. Each topic must be associated with more than one document, whose content is extracted by personalized search engines and used

to build a user profile representation.

Open Directory Project [7] (ODP) is a multi-language directory of links belonging to the Web. ODP has a hierarchic structure: the links are grouped into categories and subcategories, also known as *topics*. It is therefore possible to identify a level-based organization within the hierarchy. An example of topic is *Top/Business/Forestry and Agriculture/Fencing*; excluding the *Top* level common to all the topics, we have:

- Level I: Business;
- Level II: Forestry and Agriculture;
- Level III: Fencing.

Given the large quantity of links contained in ODP, we have decided to limit to the third level the links taken into consideration for the evaluation. The pages corresponding to such links are retrieved from the Web and indexed. The obtained index consists of 131,394 links belonging to 5,888 topics. Thereafter, ten topics are randomly chosen, five of which corresponding to potential user information needs, and five whose function is exclusively that of representing the pages visited by the user whose content is not relevant, that is, transient needs. The links of each topic were then subdivided into a training set, corresponding to 25% of the links, and the remaining links for test sets. The ten topics are summarized in Table 2.

It is clear now that this methodology allows us to build several different profiles of potential users. Once these profiles are built, it is possible to compare the precision of the search engines. In this evaluation, *Google*, *RF* and *SocialSearch* approaches are compared in terms of F1 score (or F-measure), a standard statistical measure that combines both the precision and the recall of the test to compute the resulting score. A query is built for each topic belonging to the user needs. The query is composed by the terms that form the topic name in ODP (e.g., query: "shopping craft papers"). The evaluation aims at measuring the fraction of document retrieved by the search engine from the whole collection of indexed documents that are also included in the test set for each need. Table 3 shows the variation of F1 score for the three engines. In this evaluation, *RF* engine does not take any sensible advantage of the content extracted from the training documents. *SocialSearch* outperforms the other approaches, even though several links in the training set do not have any reference in the sites of social bookmarking services. Part of the training documents are indeed very old or not very popular, therefore users are not likely to attach metadata to these resources.

[7] www.dmoz.org

5.3 WordNet-based Disambiguation Analysis

Personalization has an important role when users submit ambiguous queries, that is, consisting of terms with multiple different meanings. Past and current contexts might help disambiguate polysemous words and improve result accuracy. For this reason, this evaluation aims at gathering ambiguous queries and performs a comparison of how the different approaches behave in correctly disambiguating their meanings.

A straightforward methodology that involves the analysis of the WordNet [8] lexical database has been defined. Briefly, WordNet is a collection of *synsets*, namely, groups of nouns, adjectives and adverbs all expressing a common concept (e.g., house, home, dwelling, habitation, etc.) (Fellbaum, 1998). Synsets are interlinked by means of semantic and lexical relations. In this way, it is very easy to find terms that have potentially several different meanings (Hirst and Budanitsky, 2005).
A random choice of these ambiguous terms enables us to focus on the 12 keywords shown in Table 4. For each term, two synsets (or semantic contexts) are identified by the ones included in the database. Formally, it is possible to define a triple as follows:

$$< T, X_T, Y_T >$$

where T is a polysemic term with different meanings depending on its context, for example, $T = draft$; X_T and Y_T are two sets of tags, each of which consists of five tags that briefly describe a semantic context, for example

$$X_T = \{beer, kegerator, homebrew, keg, brewing\}$$

and

$$Y_T = \{nba\_draft, basketball, nbadraft, nba, basket\};$$

of course, T gets a different meaning in each of the two semantic contexts X and Y.

Table 4 summarizes the set of triples used in this evaluation.
For every triple, we collected 400 documents from the Web, subdivided into:

- 100 documents for each of the two contexts X and Y. These two collections are divided into two parts of 50 documents each one, which we used for training and test;
- 200 "noisy" documents, namely, that belong to both of the two semantic classes;

[8] wordnet.princeton.edu

Table 2: Benchmark statistics: ODP topic, number of links for test and training, and if topic is part of user needs.

Topic	Test links	Training links	Need
Sports/Cycling/Human Powered Vehicles	15	5	+
Computers/Home Automation/Products and Manufacturers	27	7	+
Business/Mining and Drilling/Consulting	74	18	+
Games/Roleplaying/Developers and Publishers	52	14	+
Business/Agriculture and Forestry/Fencing	100	27	+
Shopping/Crafts/Paper	35	7	
Arts/Performing Arts/Magic	25	6	
Science/Publications/Magazines and E-zines	26	7	
Science/Social Sciences/Linguistics	13	5	
Recreation/Guns/Reloading	15	5	
	382	101	

Table 3: Comparison in terms of F1 score.

Topic	PersGoogle	RF	SocialSearch
Computers/Home Automation/Products and Manufacturers	0.05	0.08	0.16
Sports/Cycling/Human Powered Vehicles	0.09	0.13	0.09
Games/Roleplaying/Developers and Publishers	0.10	0.18	0.18
Business/Mining and Drilling/Consulting	0.19	0.14	0.19
Business/Agriculture and Forestry/Fencing	0.05	0.14	0.57
Average F1	0.10	0.13	0.24

Table 4: Terms and semantic contexts.

Term	**Tags A context**	**Tags B context**
amazon	[geography, south_america, rivers, cruise, river]	[shop, books, bargains, shopping, deals]
cancer	[horoscopes, tarot, zodiac, horoscope, astrology]	[medical, medicine, health, disease, research]
capital	[dc, washington, washingtondc, washington_dc, capitolhill]	[marxism, communism, economics, socialism marx]
depression	[neuroscience, mentalhealth, psychology, health, science]	[recession, financialcrisis, economy, imf, thegreatdepression]
draft	[beer, kegerator, homebrew, keg, brewing]	[nba_draft, basketball, nbadraft, nba, basket]
hamilton	[lewishamilton, formula1, racing, mclaren, f1]	[urban canada, canadian, ontario, city]
harrison	[film, harrisonford, indianajones, movies, ford]	[george, beatles, guitar, rock, the_beatles]
lee	[kungfu, brucelee, martial, martialarts, karate]	[wii, wiimote, interaction, interface, multitouch]
mercury	[msn, java, chat, im, linux]	[planets, nasa, solarsystem, space, astronomy]
oxford	[elearning, university, courses, education, academic]	[words, language, dictionary, english reference]
porter	[5forces, marketing, strategy, management, business]	[stemmer, programming, stemming, algorithms, language]
victoria	[guide, melbourne, australia, tourism, travel]	[waterfall, safari, zambia, falls, africa]

so obtaining a collection of 4800 documents. The documents of the three collections related to the two contexts X and Y and the noisy collection are retrieved by submitting to *Delicious* the following queries, respectively:

- `q: T (tag:x1 OR tag:x2 ... tag:x5) -tag:y1 -tag:y2 ... -tag:y5`
- `q: T -tag:x1 -tag:x2 ... -tag:x5 (tag:y1 OR tag:y2 ... tag:y5)`
- `q: T -tag:x1 ... -tag:x5 -tag:y1 ... -tag:y5`

Each document retrieved by *Delicious* is annotated with a set of tags. As might be expected, two profiles U_X and U_Y are built by analyzing the documents and tags of the context X and Y, respectively. Also the noisy collection is included in both the profiles.

At the end of the training phase, the initial terms T are submitted to the search engines. For each term, the following measures are evaluated for both the profiles U_X and U_Y: P precision , R recall and $F1$ F-measure. Table 5 summarizes the results. While the order of the topics that obtain better results are similar among the considered approaches, the average precision and recall measures differ significantly. Topics such as *draft*, *mercury*, *lee* and *amazon* are clearly easier to disambiguate while *cancer*, *capital* and *depression* need more sophisticated approaches. The average precision favors *SocialSearch* and the approach based on co-occurrences. In terms of average recall and F1 score *SocialSearch* outperforms both *RF* and *CoOcc*. In particular, the average of the two standard deviation measures of F1 score over the contexts *a* and *b* shows that *SocialSearch* is able to disambiguate the same term over both the considered contexts, while *CoOcc* obtains more dispersion from the average precision.

Table 5: Average measures over all topics: Precision, Recall, F1, and its Standard Deviation.

	RF	CoOcc	SocialSearch
Avg P	0.5	0.58	0.60
Avg R	0.41	0.45	0.50
Avg F1	0.39	0.44	0.51
σ_{F1}	0.26	0.29	0.23

6 CONCLUSIONS

User generated content represents a unique source of information that can be exploited for different purposes. In this paper we have described a novel approach that takes advantage of social bookmarking services for enhancing classic query expansion techniques. Specifically, we rely on web sites such as *Delicious*, *Digg* and *StumbleUpon* in order to define semantic classes, namely, categories of terms sharing the same semantic property, based on which to categorize multiple occurrences of the query expansion process. The results of a comparative experimental evaluation confirm that the proposed approach makes for a stronger correlation among expansion terms and real user interests, thereby providing an effective solution to deal with term ambiguity.

This study shows the benefits of categorizing user interests in semantic classes related to the folksonomy extracted from social bookmarking services. As future work we plan to devise ways of integrating more social knowledge, such as social structures, in our approach. A further research effort will also concern the use of natural language processing techniques to better classify the user interests in semantic classes.

REFERENCES

Abouenour, L., Bouzouba, K., and Rosso, P. (2010). An evaluated semantic query expansion and structure-based approach for enhancing arabic question/answering. *International Journal on Information and Communication Technologies*, 3(3):37–51.

Acampora, G., Loia, V., and Gaeta, M. (2010). Exploring e-learning knowledge through ontological memetic agents. *IEEE Comp. Int. Mag.*, 5(2):66–77.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 688–695.

Biancalana, C., Flamini, A., Gasparetti, F., Micarelli, A., Millevolte, S., and Sansonetti, G. (2011). Enhancing traditional local search recommendations with context-awareness. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, UMAP'11, pages 335–340, Berlin, Heidelberg. Springer-Verlag.

Biancalana, C., Gasparetti, F., Micarelli, A., and Sansonetti, G. (2012). Enhancing query expansion through folksonomies and semantic classes. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 611–616.

Biancalana, C., Gasparetti, F., Micarelli, A., and Sansonetti, G. (2013a). An approach to social recommendation for context-aware mobile services. *ACM Trans. Intell. Syst. Technol.*, 4(1):10:1–10:31.

Biancalana, C., Gasparetti, F., Micarelli, A., and Sansonetti, G. (2013b). Social semantic query expansion. *ACM Trans. Intell. Syst. Technol.*, 4(4):60:1–60:43.

Biancalana, C., Lapolla, A., and Micarelli, A. (2009). Personalized web search using correlation matrix for query expansion. In Cordeiro, J., Hammoudi, S., and Filipe, J., editors, *WEBIST (Selected Papers)*, volume 18 of *Lecture Notes in Business Information Processing*, pages 186–198. Springer.

Biancalana, C. and Micarelli, A. (2009). Social tagging in query expansion: A new way for personalized web search. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, pages 1060–1065.

Bilotti, M. W., Katz, B., and Lin, J. (2004). What works better for question answering: Stemming or morphological query expansion? In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, Sheffield, England. ACM SIGIR.

Burgess, C. and Lund, K. (1995). Hyperspace analog to language (hal): A general model of semantic representation. In *Proceedings of the annual meeting of the Psychonomic Society*.

Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st Annual Interna-*

tional ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pages 243–250, New York, NY, USA. ACM.

Carman, M. J., Baillie, M., Gwadera, R., and Crestani, F. (2009). A statistical comparison of tag and query logs. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 123–130, New York, NY, USA. ACM.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y. (2003). Query expansion by mining user logs. *IEEE Trans. on Knowl. and Data Eng.*, 15:829–839.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Gasparetti, F. and Micarelli, A. (2003). Adaptive web search based on a colony of cooperative distributed agents. In Klusch, M., Ossowski, S., Omicini, A., and Laamanen, H., editors, *Cooperative Information Agents*, volume 2782, pages 168–183. Springer-Verlag.

Gasparetti, F., Micarelli, A., and Sansonetti, G. (2014). Exploiting web browsing activities for user needs identification. In *Proceedings of the 2014 International Conference on Computational Science and Computational Intelligence (CSCI'14)*. IEEE Computer Society, Conference Publishing Services.

Gentili, G., Marinilli, M., Micarelli, A., and Sciarrone, F. (2001). Text categorization in an intelligent agent for filtering information on the web. *IJPRAI*, 15(3):527–549.

Hirst, G. and Budanitsky, A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Nat. Lang. Eng.*, 11(1):87–111.

Jansen, B. J., Spink, A., and Pedersen, J. (2005). A temporal comparison of altavista web searching: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(6):559–570.

Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227.

Lops, P., De Gemmis, M., and Semeraro, G. (2007). Improving social filtering techniques through wordnet-based user profiles. In *Proceedings of the 11th International Conference on User Modeling*, UM '07, pages 268–277, Berlin, Heidelberg. Springer-Verlag.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Micarelli, A., Gasparetti, F., and Biancalana, C. (2006). Intelligent search on the internet. In Stock, O. and Schaerf, M., editors, *Reasoning, Action and Interaction in AI Theories and Systems*, volume 4155 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg, and New York.

Musto, C., Narducci, F., Lops, P., De Gemmis, M., and Semeraro, G. (2009). Content-based personalization services integrating folksonomies. In *Proceedings of the 10th International Conference on E-Commerce and Web Technologies*, EC-Web 2009, pages 217–228, Berlin, Heidelberg. Springer-Verlag.

Nanba, H. (2007). Query expansion using an automatically constructed thesaurus. In Kando, N. and Evans, D. K., editors, *Proceedings of the Sixth NTCIR Workshop*, pages 414–419, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan. National Institute of Informatics.

Peat, H. J. and Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42:378–383.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Salton, G. and Buckley, C. (1997). *Readings in information retrieval*, chapter Improving retrieval performance by relevance feedback, pages 355–364. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Xu, Y., Jones, G. J., and Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 59–66, New York, NY, USA. ACM.

Zamir, O. E., Korn, J. L., Fikes, A. B., and Lawrence, S. R. (2004). Us patent application #0050240580: Personalization of placed content ordering in search results.

Zhou, D., Lawless, S., and Wade, V. (2012). Improving search via personalized query expansion using social media. *Inf. Retr.*, 15(3-4):218–242.

An Overview of Cloud Identity Management-Models

Bernd Zwattendorfer, Thomas Zefferer and Klaus Stranacher
Institute for Applied Information Processing and Communications (IAIK), Graz University of Technology, Inffeldgasse 16a, 8010 Graz, Austria
{*bernd.zwattendorfer, thomas.zefferer, klaus.stranacher*}*@iaik.tugraz.at*

Keywords: Cloud, Cloud Computing, Cloud Identity Management-Model, Identity Management.

Abstract: Unique identification and secure authentication are essential processes in various areas of application, e.g. in e-Government, e-Health, or e-Business. During the past years several identity management-systems and models have evolved. Many organizations and enterprises or even countries for their national eID solutions rely on identity management-systems for securing their applications. Since more and more applications are migrated into the cloud, secure identification and authentication are also vital in the cloud domain. However, cloud identity management-systems need to meet slightly different requirements than traditional identity management-systems and thus cannot be clustered into the same model types or categories. Therefore, in this paper we give an overview of different cloud identity management-models that have already emerged up to now. We further compare these models based on selected criteria, e.g. on practicability and privacy aspects.

1 INTRODUCTION

Secure and reliable identity management (IdM) plays a vital role in several security-sensitive areas of applications, e.g. in e-Government, e-Business, or e-Health. An identity management-system helps online applications to control access for users to protected resources or services. However, identity management is no new topic and several identity management-approaches and systems have already emerged over time. A comprehensive overview on identity management-systems is given in (Bauer et al., 2005).

Due to the increasing number of cloud computing adoption and the deployment of security-sensitive cloud applications, secure identity management becomes also more and more important in the cloud domain. In addition, outsourcing identity management-systems to the cloud can bring up several benefits such as higher scalability or cost savings, since no in-house infrastructure needs to be hosted and maintained. However, the field of cloud identity management is still new and not extensively investigated yet. Therefore, the aim of this paper is to overview different cloud identity management-models, discuss advantages and disadvantages of the individual models, provide a comprehensive survey, and finally compare them based on selected criteria. The criteria for the comparison have been selected by focusing on practicability and privacy, since one of the main issues of cloud computing is the loss of data protection and privacy (Pearson and Benameur, 2010), (Zissis and Lekkas, 2012), and (Sen, 2013).

The paper is structured as follows. Section 2 classifies existing traditional identity management-models and their implementations. Section 3 surveys existing cloud identity management-models and describes their benefits and drawbacks. These models are compared in Section 4 based on selected criteria. Finally, conclusions are drawn in Section 5.

2 TRADITIONAL IDENTITY MANAGEMENT-MODELS

An identity management-system usually involves four entities (Bertino and Takahashi, 2011). A *service provider* (SP) provides different online services to *users*. Before being allowed to consume such services, a user has to successfully identify and authenticate. Therefore, the user usually identifies and authenticates at a so-called *identity provider* (IdP). The identity provider is then in charge of providing the users identity data and supplementary authentication results to the service provider in a secure way. Finally, a *control party*, which is usually a law or regulation enforcing body, needs to investigate identity data transactions, e.g. for data protection reasons. Hence,

main purpose of such control party is auditing. Figure 1 illustrates the communication process in an identity management-system including all four entities.

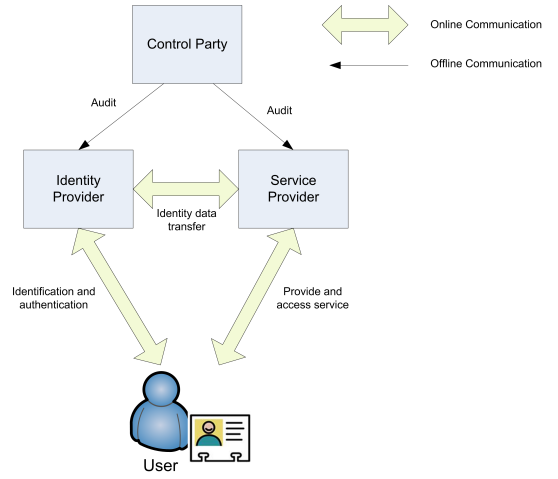


Figure 1: Entities involved in an identity management-system.

Over time, several identity models involving these four entities and supporting similar but slightly different use cases have evolved. Some of these models have advantages in scalability, others in privacy or user control. In the following subsections we briefly describe the most important models based on the work of (Cao and Yang, 2010), (Dabrowski and Pacyna, 2008), (Dbrowski and Pacyna, 2008), (Jøsang et al., 2005), (Jøsang and Pope, 2005), (Jøsang et al., 2007), and (Palfrey and Gasser, 2007). For simplicity, we skip a discussion of the control party in all subsequent models because its functionality remains the same in all models.

2.1 Isolated Model

The *isolated model* is basically the simplest traditional identity model. In this model, the service provider and identity provider merge, hence identification and authentication are directly carried out at the service provider. In addition, the functionality of the identity management-system (creating, maintaining, or deleting identities) can only be used by this specific service provider. If a user wants to access services of another service provider, she needs to register at the other service providers identity management-system again. This further means that each individual service provider has to store and maintain the identity data and credentials of the user separately. While this still may not be a huge burden for service providers, the diversity of credentials for accessing various service providers may become unmanageable for users (Jøsang and Pope, 2005). This model can still be found by service providers on the Internet.

2.2 Central Model

The *central identity model* avoids diverse identity management-systems, where the user has to register separately. Instead, the identity management-system is outsourced by several service providers to a central identity provider. The identity provider takes over all identity-related functionality for the service provider, including credential issuance, identification and authentication, and the management of the identity lifecycle in general (Bertino and Takahashi, 2011). Furthermore, in this model users' identity data are stored in a central repository at the identity provider and service providers do not need to maintain identity data in their own repositories (Cao and Yang, 2010). For authentication at a service provider, the user has to identify and authenticate at the identity provider before. The identity provider then assembles a token including all necessary identity and authentication information of the user and transmits it to the service provider[1]. (Jøsang et al., 2005) further distinguish the domain model for the identifier used. In the *common identifier model* one and the same identifier is used for identification at all service providers. In contrast to that, in the *meta identifier domain model* separate identifiers are used for identification at the individual service providers. However, all separate identifiers map to a common meta identifier at the identity provider to uniquely identify the user. Typical examples implementing this approach are Kerberos (Neuman et al., 2005) or the Central Authentication Service (CAS)[2].

2.3 User-centric Model

While in the central model all identity data of the user are stored in the domain of the identity provider, in the *user-centric model* all identity data are stored directly in the users domain, e.g. on a secure token such as a smart card. The main advantage of this model is that the user always remains the owner of her identity data and stays under their full control (Dbrowski and Pacyna, 2008). Identity data can only be transferred by an identity provider to a service provider if the user explicitly gives her consent to do so. Compared to the central model, this tremendously increases users' privacy. (Jøsang and Pope, 2005) discuss in detail this user-centric approach. Typical examples imple-

[1] Different approaches exist; hence identity data can be either pushed to or pulled from the service provider.

[2] http://www.jasig.org/cas

menting this model are Windows CardSpace[3] or various national eID solutions such as the Austrian citizen card (Leitold et al., 2002) or the German eID (Frommm and Hoepner, 2011).

2.4 Federated Model

In the *federated model* identity data are not stored in a central repository but are rather stored distributed across different identity and/or service providers. No single entity is fully controlling the identity information (Palfrey and Gasser, 2007). The distributed identity data of a particular user are linked usually by the help of a common identifier[4]. All identity providers and service providers, which take part in such a federation, share a common trust relationship amongst each other. The trust relationship is usually established on organizational level whereas enforcement is carried out on technical level. This federated model particularly supports identification and authentication across different domains, which paves the way for cross-domain single sign-on (Cao and Yang, 2010). Popular examples of this approach are the Security Assertion Markup Language (SAML)[5], Shibboleth[6], or WS-Federation (Kaler and McIntosh, 2009).

3 CLOUD IDENTITY MANAGEMENT-MODELS

Cloud computing is currently still one of the most emerging trends in the IT sector. Many applications are already migrated to the cloud because of its benefits such as cost savings, scalability, or less maintenance efforts (Armbrust et al., 2009). Due to the increasing number of cloud applications, secure identity management is equally important for cloud applications as for traditional web applications. Hence, new cloud identity management-models have already emerged, which particularly take the properties of cloud computing into account. (Cloud Security Alliance, 2011), (Cox, 2012), (Gopalakrishnan, 2009), (Goulding, 2010), or (Zwattendorfer et al., 2013) already describe cloud identity management-models in their publications. We take these publications as a basis to give an overview of different existing cloud identity management-models. In the following subsections we describe the individual models in more detail and explain how and where identities are stored and managed.

3.1 Identity in the Cloud-Model

The *Identity in the Cloud-Model* is similar to the *isolated identity model* described in Section 2.1. Again, identity provider and service provider merge also in this model. This means for the cloud case that the cloud service provider, which hosts the application, is also responsible for the identity management. Figure 2 illustrates this model.

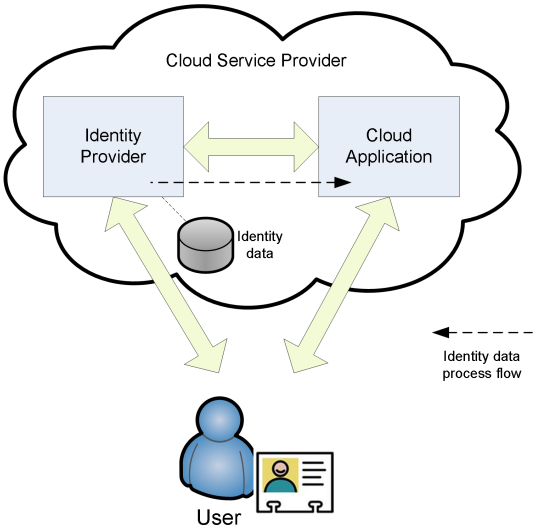


Figure 2: Identity in the Cloud-Model.

Identity data of users, who are accessing the cloud application, are directly stored in the domain of the cloud service provider. Hence, the user has actually no control which data are processed in the cloud. Cloud service providers which already use this model for their Software as a Service (SaaS) applications are for instance Google or Salesforce.com. They offer their own user management to their customers for managing their own identities. The main advantage of this model is that organizations do not need to host and maintain their own identity management-system but can simply rely on an existing one, which will be maintained by the cloud service provider. Needless to say that costs can be decreased at an organization when applying this model. However, the use of this model also shifts responsibility in terms of security and privacy to the cloud service provider and the organization more or less looses control over the identity data stored and managed in the cloud.

[3]http://msdn.microsoft.com/en-us/library/vstudio/ms733090%28vvs.90%29.aspx

[4]It is not necessary that the common identifier is shared. Different identifiers mapping to the same user are also possible (Cao and Yang, 2010).

[5]http://saml.xml.org

[6]http://shibboleth.net

3.2 Identity to the Cloud-Model

The *Identity to the Cloud-Model* is similar to the traditional *central identity model*. Also in this model, the identity provider takes over the tasks regarding identity management for the service provider. However, the main difference in this model is that the service provider and its applications are cloud-based. This further means that in this model the identity provider is not deployed in the cloud, which avoids unnecessary identity data disclosure to a cloud service provider. Figure 3 illustrates the *Identity to the Cloud-Model*.

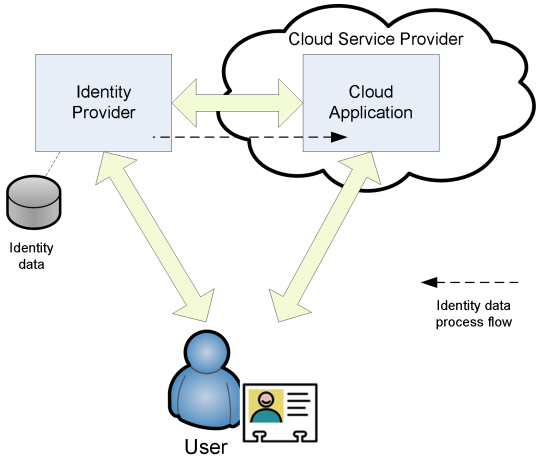


Figure 3: Identity to the Cloud-Model.

In more detail, the complete user and identity management is still hosted by the organization e.g. in one of its data centers. Before gaining access to a cloud application, users have to authenticate at the identity provider first. After that, the identity provider transfers appropriate identity and authentication data to the cloud service provider through well-defined and standardized interfaces. Google or Salesforce.com, for instance, rely on SAML, OpenID[7], or OAuth[8] for these interfaces and external identity provisioning.

Appliance of this model has the advantage that an existing identity management-infrastructure of an organization can be re-used. Users are identified and authenticated at the cloud application by the use of this external identity management-system. No new user management has to be created or migrated to the cloud service provider. The organization remains under control of the identity data and provides it to the cloud service provider just on demand. However, interoperability issues may arise due to the use of external interfaces. For instance, a common agreement on the attributes transferred (e.g. format or semantic) between the identity provider and the cloud service provider must be given. In addition, the identity provider must support the interface provided by the cloud service provider.

3.3 Identity from the Cloud-Model

The *Identity from the Cloud-Model* fully features the cloud computing paradigm. In this case, both the cloud application and the identity provider are operated in the cloud. However, in contrast to the *Identity in the Cloud-Model* of Section 3.1 both entities are operated by distinct cloud service providers. Since identities are provided as a service from the cloud, this model is also named "Identity as a Service Model" (Ates et al., 2011). Google or Facebook are for instance such providers, when using the authentication functionality for other services than their own (Google Accounts Authentication and Authorization[9] or Facebook Login[10]). Figure 4 illustrates this model.

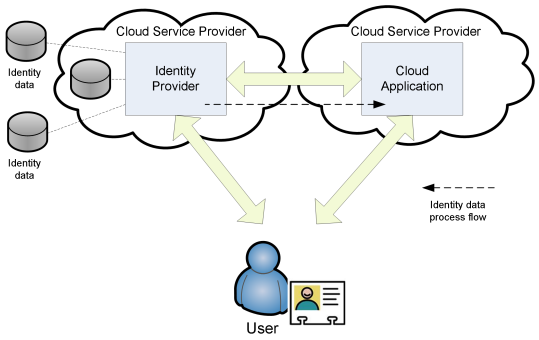


Figure 4: Identity from the Cloud-Model.

By applying this model, an organization can benefit from the pure cloud computing advantages such as high scalability or elasticity. Besides that, compared to the previous cloud identity management-models the advantage of this model is the separation of cloud service providers. In this model, organizations can select their preferred identity provider in the cloud. This is particularly important because the organization needs to trust the identity provider, which is responsible for the organization's identity and user management. Organizations must be careful in cloud service provider selection, as e.g. legal implications such as data protection regulations might hinder the selection of a provider which stores identity data in a foreign country.

[7] http://openid.net
[8] http://oauth.net
[9] https://developers.google.com/accounts
[10] https://developers.facebook.com/docs/facebook-login

3.4 Cloud Identity Broker-Model

The *Cloud Identity Broker-Model* can be seen as an extension to the *Identity from the Cloud-Model*. In this *Cloud Identity Broker-Model*, the identity provider in the cloud acts now as an identity broker in the cloud. In other words, the cloud identity broker is some kind of hub between one or more service providers and one or more identity providers. Figure 5 illustrates this model.

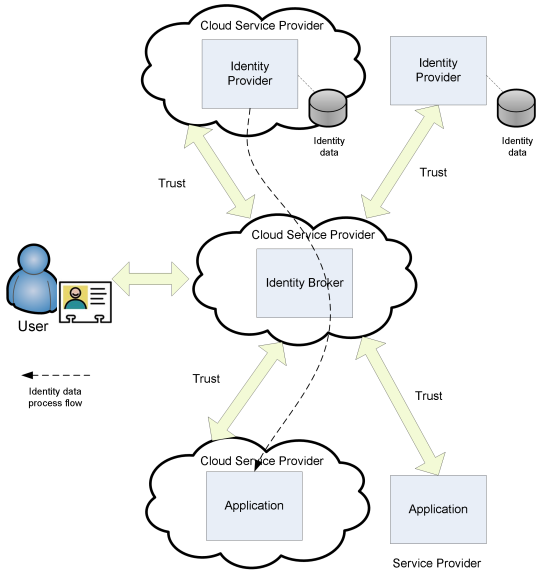


Figure 5: Cloud Identity Broker-Model.

The basic idea behind this model is to decouple the service provider from integrating and connecting a vast amount of identity providers. If no broker is used, a single service provider has to implement all interfaces for communication with the individual identity providers if the service provider wants to support them. By applying the broker concept, the identity broker hides the complexity of the individual identity providers from the service provider. This further means that the service provider just needs to implement one interface, namely the one to the identity broker. All other interfaces are encapsulated by the identity broker and tailored or mapped to the service provider's interface. In addition, for the service provider only one strong trust relationship between the service provider and the identity broker is required. All other trust relationships with the individual identity providers are "brokered" by the identity broker. Deploying the broker in the cloud makes this model even more powerful. Due to the cloud advantages of nearly unlimited computing resources and scalability, a high number of active connections and identification/authentication processes at the broker can be easily absorbed by the cloud.

Nevertheless, still some disadvantages can be found in this model. One disadvantage is that both the user and the service provider are dependent on the functionality the cloud identity broker supports. If the identity broker does not support the desired identity provider the user wants to use for authentication, the service provider cannot provide its services to the user. Furthermore, if the broker does not support the communication interface to the service provider anymore, the service provider is cut off from any other identity provider. However, probably the main issue is that identity data runs through the cloud identity broker in plaintext. As already mentioned before, privacy issues concerning the cloud service provider might hinder adoption of this cloud-based identity management-service (Pearson and Benameur, 2010).

The *Cloud Identity Broker-Model* has already been implemented by some organizations. McAfee Cloud Single Sign On[11], the SkIDentity[12] implementation, or the Cloud ID Broker[13] of Fugen are just a few examples. Further details on this model can be found in (Cloud Security Alliance, 2011), (Huang et al., 2010), or (Zwattendorfer et al., 2013).

3.5 Federated Cloud Identity Broker-Model

The *Federated Cloud Identity Broker-Model* combines the traditional *federated identity model* with the newly *Cloud Identity Broker-Model*. This combined model has been introduced by (Zwattendorfer et al., 2013) and aims on eliminating the drawbacks of the central *Cloud Identity Broker-Model*. The general architecture is illustrated in Figure 6, showing the federation of two different cloud identity brokers.

Compared to the simple *Cloud Identity Broker-Model*, in this federated model users and service providers do not need to rely on one and the same identity broker. Actually, both the user and the service provider can rely on the individual broker of their choice. This eliminates the drawback for both the user and the service provider of being dependent on the same identity broker. On the one hand, users can simply select the identity broker that supports all their desired identity providers (Identity Broker 1 in Figure 6). On the other hand, service providers can select the broker that e.g. supports a specific communication interface (Identity Broker 2 in Figure 6). Hence, referring to Figure 6 the communication process flow

[11] http://www.mcafee.com/us/products/cloud-single-sign-on.aspx

[12] http://www.skidentity.com

[13] http://fugensolutions.com/cloud-id-broker.html

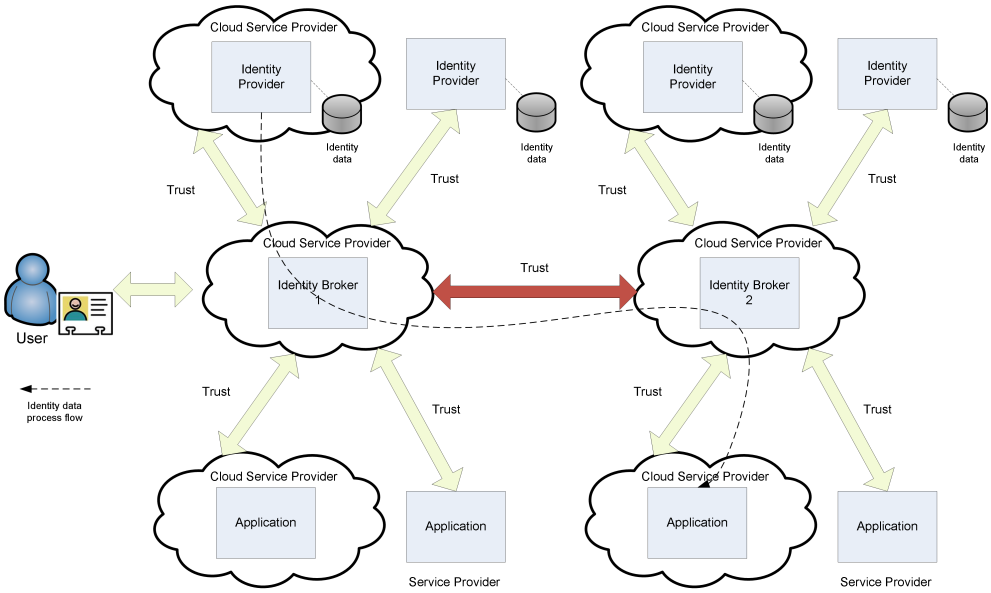


Figure 6: Federated Cloud Identity Broker-Model.

between identity provider and service provider is brokered through the two Identity Brokers 1 and 2.

While this model eliminates some problems of the *Cloud Identity Broker-Model*, the issue of plain identity data transfer between and through cloud service providers still persists. To bypass such privacy issue, the following two models had been introduced.

3.6 BlindIdM-Model

The *BlindIdM-Model* has been introduced by (Nuñez et al., 2013) and (Nuñez and Agudo, ress)[14] and can also be seen as an extension and alteration of the *Identity from the Cloud-Model*. The basic idea is principally the same, however, this model enables identity data storage and data processing also by semi-trusted identity providers[15] in the cloud. In fact, the identity provider in the cloud can provide identity data to service providers without actually knowing the contents of these data. Hence, the identity provider provides these data in a blind manner (Nuñez et al., 2013). This particularly preserves users' privacy, as only blinded data is transferred through the cloud identity provider and the cloud provider has no possibility to inspect these data.

The identity data being transferred are actually blinded by using a proxy re-encryption scheme[16] (Green and Ateniese, 2007) (Ateniese et al., 2006). In more detail, during identity management setup and user registration the organization stores the users' identity data in encrypted format at the cloud identity provider. Thereby, the private key is kept confidential by the organization, hence the cloud provider is not able to decrypt the stored identity data. In addition, the organization generates a re-encryption key for the identity provider[17], which allows the re-encryption from the stored data encrypted for the cloud identity provider into other encrypted data, which however can be decrypted by the service provider. During an authentication process, the cloud identity provider then just re-encrypts the desired identity data of the user for the service provider. The practical applicability of the *BlindIdM-Model* has been shown by an implementation in connection with OpenID (Nuñez et al., 2012).

3.7 Privacy-Preserving Federated Cloud Identity Broker-Model

The main aim of this model is – similar to the *BlindIdM-Model* – an improved privacy-preservation for the user. Thereby, the same concept of "blinding"

[14] a similar approach has been introduced by (Zwattendorfer and Slamanig, 2013a)

[15] A semi-trusted identity provider is an identity provider that works correctly but may be interested in inspecting private data. In other words, the identity provider acts *honest but curious*.

[16] By using proxy re-encryption a semi-trusted proxy can alter a ciphertext, which has been encrypted for person A, in such a way that it can be decrypted by person B. Thereby, the proxy gains no access to the plaintext of the data.

[17] For generating a re-encryption key, the organization requires its private key and the public key of the service provider.

identity data is applied to the basic *Federated Cloud Identity Broker-Model*. Hence, this model combines the advantages of the *Federated Cloud Identity Broker-Model* with the advantages of the *BlindIdM-Model*. Furthermore, this model can again be applied when having semi-trusted cloud identity brokers. The general concept of this model has been introduced by (Zwattendorfer, 2014).

The general concept of this model is similar to the *BlindIdM-Model* because also proxy re-encryption is used for protecting identity data from the cloud service providers. However, the main differences are that the data can also be stored encrypted at non-cloud identity providers and that the data can also be encrypted by the user and not only by an organization. In addition – which is the basic concept of this federated model – there are two re-encryption steps required, since identity data needs to flow at least through two cloud identity brokers. For instance, lets assume that the user has stored some identity data, which are encrypted for Identity Broker 1, at an identity provider. To successfully run such a privacy-preserving authentication process, the user additionally has to generate two re-encryption keys (One for the direction Identity Broker 1 $\rightarrow$ Identity Broker 2 and one for the direction Identity Broker 2 $\rightarrow$ service provider) and issues them to the respective entities. Finally, after successful authentication at the identity provider, identity data are transferred through the chain identity provider $\rightarrow$ Identity Broker 1 $\rightarrow$ Identity Broker 2 $\rightarrow$ service provider by applying proxy re-encryption in the last two steps (The identity data was already encrypted for Identity Broker 1 during storage at the identity provider, hence only two instead of three re-encryption steps are required). An application of this model can be found in (Zwattendorfer and Slamanig, 2013b), where parts of the STORK[18] framework are realized using this architecture to enhance scalability by ensuring users' privacy at the same time.

4 COMPARISON OF CLOUD IDENTITY MANAGEMENT-MODELS

In this section we evaluate, discuss, and compare the various cloud identity management-models based on different criteria. Comparison criteria are defined in the following Subsection 4.1 whereas the comparison itself is elaborated in Subsection 4.2.

[18]Secure Identity Across Borders Linked, https://www.eid-stork.eu/

4.1 Comparison Criteria

The following criteria act as a basis for comparing the various cloud identity management-models. Some of the comparison criteria were selected or derived from (Cao and Yang, 2010), (Nuñez et al., 2013), and (Birrell and Schneider, 2013). The selected criteria target aspects of different areas (e.g. general architecture, trust, privacy, etc.). The diversity of the criteria was deliberately considered to give a comprehensive overview on the different cloud identity management-models.

Number of SPs Supported: Is the model limited to one SP or can multiple SPs be supported?

Number of IdPs Supported: Is the model limited to one IdP or can multiple IdPs be supported?

Trust Domains: Is authentication supported only within a single trust domain or also across different trust domains?

Trust Model: Is a direct trust model or a brokered trust model applied?

Trust in the Cloud IdP/identity Broker: Must the cloud identity provider/identity broker be trusted or can they be semi-trusted?

Single sign-on (SSO): Can the model support single sign-on (SSO)?

Storage Location of Identity Data: Where are users' identity data stored?

Scalability: Is the model applicable in a large scale?

Extensibility: Is the model easily extensible, e.g. by adding new service providers?

Governance Framework: Is a governance framework involving several entities required?

Cost Effectiveness: Is the model cost effective?

Confidentiality: Does the identity data stay confidential at the identity provider/identity broker?

Minimal/Selective Disclosure: Can the user select the amount of identity data to be disclosed to the identity provider/service provider?

User Control: Does the user have full control over her identity data?

Unlinkability: Is the user unlinkable to the identity provider/identity broker? In other words, are different authentication processes of the same user linkable?

Anonymity: Can the user stay anonymous with respect to the identity provider/identity broker?

4.2 Comparison

In this section we compare the individual cloud identity management-models with respect to the prior defined criteria. Table 1 shows and summarizes this comparison. For some comparisons we use qualitative arguments, for others quantitative arguments (low, medium, high), and for the rest simply boolean (e.g. yes/no for being applicable or not) arguments. The options marked in bolt indicate the respective best option (only applicable for quantitative and boolean values). The underlying principle for all comparisons (in particular for those that are related to privacy such as confidentiality, minimal/selective disclosure, etc.) is that we assume an identity provider or an identity broker deployed in the cloud acting *honest but curious* (thus being semi-trusted). In contrast to that we assume applications in the cloud and their hosting service providers as being trusted, as they anyhow require users' identity data for service provisioning.

In the following we discuss the various models based on the individual criteria.

Number of SPs Supported: Since in the *Identity in the Cloud-Model* the service provider and the identity provider are the same entity, the identity provider can only serve one service provider. All other models have no such restriction and thus can provide multiple service providers with identity data.

Number of IdPs Supported: Only those models that rely on a broker-based approach are able to deal with multiple connected identity providers. All others just include one identity provider. Dealing with multiple identity providers has the advantage that a user can simply select her preferred identity provider for an authentication process. Different identity providers can have different identity data stored or support different qualities in the authentication mechanisms. This allows users to select the identity provider satisfying best the needs for authentication at a service provider.

Trust Domains: The broker-based models support authentication across multiple trust domains, as multiple entities are involved during an authentication process. All others support authentication in single domains only.

Trust Model: Again, all models which rely on an identity broker also feature a brokered trust model, hence the trust relationships are segmented. All other models rely on a direct or pairwise trust model, as only the service provider and the identity provider communicate with each other during an authentication process. A clear statement which model has more advantages cannot be made. Both have their benefits and drawbacks, however, details on the individual models can be found in (Linn et al., 2004).

Trust in the Cloud IdP/identity Broker: For the two models (*BlindIdM-Model* and *Privacy-Preserving Federated Cloud Identity Broker-Model*), which rely on proxy re-encryption for securing the data during cloud transmission, it is sufficient when the identity provider/identity broker is considered semi-trusted. In all other cloud identity models the identity provider/identity broker must be trusted.

Single sign-on (SSO): In fact, all models that can handle multiple service providers are principally applicable to support single sign-on. This means, that only the *Identity in the Cloud-Model* cannot support a simplified log-in process.

Storage Location of Identity Data: In the *Identity to the Cloud-Model* identity data are stored on a single external identity provider, which is capable of providing identity to the cloud application through a well-defined interface. In the broker-based models, identity data can be stored distributed across multiple different identity providers, being either deployed in the cloud or in a conventional data center. However, the different identity providers could also have identity data stored redundantly, i.e. the same attribute name/value-pair is stored at different providers. No identity data are actually stored at the identity broker. In the remaining cloud identity models identity data are stored directly at the cloud identity provider.

Scalability: The *Identity to the Cloud-Model* has the lowest scalability, as an external identity provider is usually not designed for dealing with high load activities. In addition, an external identity provider has not that flexibility or elasticity that an identity provider deployed in a cloud has. Hence, such cloud identity providers (*Identity in the Cloud-Model*, *Identity from the Cloud-Model*, and *BlindIdM-Model*) have higher scalability features. Although in these three models the identity provider/identity broker is deployed in the cloud, we rated the models with just medium level scalability. The reason is that with the broker-based models load can additionally be distributed to other identity providers and thus is not bundled at one single provider. Hence, the broker-based models achieve the highest scalability.

Table 1: Comparison of the individual cloud identity management-models based on selected criteria.

Criterion / Model	**Identity in the Cloud-Model**	**Identity to the Cloud-Model**	**Identity from the Cloud-Model**	**Cloud Identity Broker-Model**	**Federated Cloud Identity Broker-Model**	**BlindIdM-Model**	**Privacy-Preserving Federated Cloud Identity Broker-Model**
Number of SPs supported	One	**Multiple**	**Multiple**	**Multiple**	**Multiple**	**Multiple**	**Multiple**
Number of IdPs supported	One	One	One	**Multiple**	**Multiple**	One	**Multiple**
Trust domains	One	One	One	**Multiple**	**Multiple**	One	**Multiple**
Trust model	Direct	Direct	Direct	Brokered	Brokered	Direct	Brokered
Trust in the cloud IdP/identity broker	Trusted	Trusted	Trusted	Trusted	Trusted	**Semi-Trusted**	**Semi-Trusted**
Single sign-on (SSO)	No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Storage location of identity data	Cloud identity provider	External identity provider	Cloud identity provider	Cloud identity provider and external identity provider	Cloud identity provider and external identity provider	Cloud identity provider	Cloud identity provider and external identity provider
Scalability	Medium	Low	Medium	**High**	**High**	Medium	**High**
Extensibility	Low	Medium	Medium	**High**	**High**	Medium	**High**
Governance framework	**No**	**No**	**No**	Yes	Yes	Yes	Yes
Cost effectiveness	Medium	Medium	Medium	**High**	**High**	Medium	**High**
Confidentiality	No	No	No	No	No	**Yes**	**Yes**
Minimum/Selective disclosure	No	**Yes**	No	**Yes**	**Yes**	No	**Yes**
User Control	No	**Yes**	No	**Yes**	**Yes**	No	**Yes**
Unlinkability	No	No	No	No	No	No	**Yes**
Anonymity	No	No	No	No	No	**Yes**	**Yes**

Extensibility: The *Identity in the Cloud-Model* cannot be extended because service provider and identity provider are one and the same entity. The *Identity to the Cloud-Model*, the *Identity from the Cloud-Model*, and the *BlindIdM-Model* can be extended to integrate additional service providers. Nevertheless, the broker-based models have the best extensibility as from their nature the general aim is to support multiple service providers and identity providers.

Governance Framework: The non-broker-based cloud identity models do not require an extensive governance framework as only a simple pairwise (direct) trust model applies. In the broker-based concepts a thorough governance framework is required as multiple providers have to interact. For the privacy-preserving models (*BlindIdM-Model* and *Privacy-Preserving Federated Cloud Identity Broker-Model*) the governance framework gets even more complex, as encryption keys have to be managed for the individual entities.

Cost Effectiveness: The broker-based models have the highest cost effectiveness, since the identity brokers are deployed in the cloud and additionally multiple identity providers can be connected and re-used. Due to the re-use of existing external identity providers, costs can be saved. The same arguments also hold for the *Identity to the Cloud-Model*, where an existing identity management-system through an external interface is re-used for identity data provisioning. However, this model cannot benefit from the advantages of an identity provider in the cloud deployment, which leads to medium cost effectiveness only. All other models also have medium cost effectiveness, as the identity provider is deployed in the cloud but no existing identity providers can be re-used.

Confidentiality: Only the *BlindIdM-Model* and the *Privacy-Preserving Federated Cloud Identity Broker-Model* support confidentiality with respect

to the cloud service provider because the identity data transferred through the cloud service provider are encrypted. In comparison, in all other cloud identity models identity data are routed in plaintext through the cloud service provider that hosts the cloud identity provider/identity broker.

Minimum/Selective Disclosure: For evaluating this criterion we assume that minimum/selective disclosure is only possible at trusted identity providers. Hence, this feature is only supported where external (and trusted) identity providers are part of the model. These are the broker-based models as well as the *Identity to the Cloud-Model*. All other models rely on cloud identity providers only.

User Control: Again, for evaluating this criterion we assume that full user-control is only possible at trusted identity providers. Therefore, the same results as for the comparison with respect to minimum/selective disclosure apply.

Unlinkability: The user – in fact – is only unlinkable with respect to the identity broker in the *Privacy-Preserving Federated Cloud Identity Broker-Model*. The reasons are that, on the one hand, the identity broker just sees encrypted data and, on the other hand, that the encrypted data can be randomized if certain proxy re-encryption schemes such as from (Ateniese et al., 2006) are used. The randomization feature allows to provide the identity broker with different ciphertexts during different authentication processes although the containing plaintext data remains the same. Hence, this avoids user linkage during different authentication processes of the same user. Although the *BlindIdm-Model* supports proxy re-encryption too, the randomization feature has no effect in this case because the encrypted data are directly stored at the cloud identity provider. If the user wants to update her encrypted identity data at the cloud identity provider, she must somehow be linkable. All other models also do not support unlinkability because identity data flows through the identity provider/identity broker in plaintext.

Anonymity: The only two models that support anonymity with respect to the identity broker are the *BlindIdM-Model* and the *Privacy-Preserving Federated Cloud Identity Broker-Model*. In these two models the identity data are fully hidden from the identity broker due to encryption. Even if the user is linkable, the broker cannot reveal the user's identity. In all other models anonymity with respect to the identity provider/identity broker is not possible because identity data are processed in plaintext.

5 CONCLUSIONS

Based on the comparison and discussion of the different cloud identity management-models it can be concluded that the *Privacy-Preserving Federated Cloud Identity Broker-Model* does the best with respect to the selected criteria. It supports the main basic functions like all other cloud identity models but additionally tremendously increases users' privacy. However, application of this model is also more complex than the others. Reasons are the support of authentication across several domains of multiple identity providers and service providers and the incorporation of privacy features due to the use of proxy re-encryption. Furthermore, the use of proxy re-encryption requires a thorough key management, which implies the necessity of an appropriate governance framework. In addition, the brokered trust model might be a blocking issue for further adoption of this model as liability is shifted to the intermediary components (identity brokers). However, in general the broker-based cloud identity management-models have more advantages than the simple cloud identity management-models. Nevertheless, the use of any cloud identity management-model is advantageous compared to traditional identity management-models as they provide higher scalability and better cost effectiveness due to the cloud computing features.

REFERENCES

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. (2009). Above the Clouds : A Berkeley View of Cloud Computing Cloud Computing. Technical report, RAD Lab.

Ateniese, G., Fu, K., Green, M., and Hohenberger, S. (2006). Improved proxy re-encryption schemes with applications to secure distributed storage. *ACM Trans. Inf. Syst. Secur.*, 9(1):1–30.

Ates, M., Ravet, S., Ahmat, A. M., and Fayolle, J. (2011). An Identity-Centric Internet: Identity in the Cloud, Identity as a Service and Other Delights. *ARES 2011*, pages 555–560.

Bauer, M., Meints, M., and Hansen, M. (2005). D3.1: Structured Overview on Prototypes and Concepts of Identity Management System. FIDIS.

Bertino, E. and Takahashi, K. (2011). *Identity Management: Concepts, Technologies, and Systems*. Artech House.

Birrell, E. and Schneider, F. (2013). Federated Identity Management Systems: A Privacy-based Characterization. *IEEE Security and Privacy*, 11(5):36–48.

Cao, Y. and Yang, L. (2010). A survey of Identity Management technology. In *IEEE ICITIS 2010*, pages 287–293. IEEE.

Cloud Security Alliance (2011). Security Guidance for Critical Areas of Focus in Cloud Computing V3.0. CSA.

Cox, P. (2012). How to Manage Identity in the Public Cloud. *InformationWeek reports*.

Dabrowski, M. and Pacyna, P. (2008). Generic and Complete Three-Level Identity Management Model. In *SECURWARE 2008*, pages 232–237. IEEE.

Dbrowski, M. and Pacyna, P. (2008). Overview of Identity Management. Technical report, chinacommunications.cn.

Frommm, J. and Hoepner, P. (2011). The New German eID Card. In Fumy, W. and Paeschke, M., editors, *Handbook of eID Security - Concepts, Practical Experiences, Technologies*, pages 154–166. Publicis Publishing, Erlangen.

Gopalakrishnan, A. (2009). Cloud Computing Identity Management. *SETLabs Briefings*, 7(7):45–55.

Goulding, J. T. (2010). identity and Access Management for the Cloud : CAs Strategy and vision. Technical Report May, CA Technologies.

Green, M. and Ateniese, G. (2007). Identity-Based Proxy Re-encryption. In *ACNS 2007*, volume 4521 of *LNCS*, pages 288–306. Springer.

Huang, H. Y., Wang, B., Liu, X. X., and Xu, J. M. (2010). Identity Federation Broker for Service Cloud. *ICSS 2010*, pages 115–120.

Jøsang, A., Fabre, J., Hay, B., Dalziel, J., and Pope, S. (2005). Trust Requirements in Identity Management. *Proceedings of the 2005 Australasian workshop on Grid computing and e-research*, pages 99–108.

Jøsang, A. and Pope, S. (2005). User Centric Identity Management. *AusCERT 2005*.

Jøsang, A., Zomai, M. A., and Suriadi, S. (2007). Usability and privacy in identity management architectures. In *ACSW '07*, pages 143–152.

Kaler, C. and McIntosh, M. (2009). Web Services Federation Language (WS-Federation) Version 1.2. OASIS Standard.

Leitold, H., Hollosi, A., and Posch, R. (2002). Security architecture of the Austrian citizen card concept. In *ACSAC 2002*, pages 391–400.

Linn, J., Boeyen, S., Ellison, G., Karhuluoma, N., Macgregor, W., Madsen, P., Sengodan, S., Shinkar, S., and Thompson, P. (2004). Trust Models Guidelines. Technical report, OASIS.

Neuman, C., Yu, T., Hartman, S., and Raeburn, K. (2005). The Kerberos Network Authentication Service (V5). RFC 4120 (Proposed Standard).

Nuñez, D., Agudo, I., and Lopez, J. (2013). Leveraging Privacy in Identity Management as a Service through Proxy Re-Encryption. In Zimmermann, W., editor, *Proceedings of the PhD Symposium at the 2nd European Conference on Service-Oriented and Cloud Computing*, pages 42–47.

Nuñez, D. and Agudo, I. (In Press). Blindidm: A privacy-preserving approach for identity management as a service. *International Journal of Information Security*.

Nuñez, D., Agudo, I., and Lopez, J. (2012). Integrating OpenID with Proxy Re-Encryption to enhance privacy in cloud-based identity services. In *IEEE CloudCom 2012*, pages 241 – 248.

Palfrey, J. and Gasser, U. (2007). CASE STUDY: Digital Identity Interoperability and eInnovation. Berkman Publication Series,.

Pearson, S. and Benameur, A. (2010). Privacy, Security and Trust Issues Arising from Cloud Computing. In *CloudCom 2010*, pages 693–702. IEEE.

Sen, J. (2013). Security and Privacy Issues in Cloud Computing. In Martínez, A. R., Marin-Lopez, R., and Pereniguez-Garcia, F., editors, *Architectures and Protocols for Secure Information Technology Infrastructures*, pages 1–45. IGI Global.

Zissis, D. and Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3):583–592.

Zwattendorfer, B. (2014). Towards a Privacy-Preserving Federated Identity as a Service Model. to appear.

Zwattendorfer, B. and Slamanig, D. (2013a). On Privacy-Preserving Ways to Porting the Austrian eID System to the Public Cloud. In *SEC 2013*, AICT, pages 300–314. Springer.

Zwattendorfer, B. and Slamanig, D. (2013b). Privacy-Preserving Realization of the STORK Framework in the Public Cloud. In *SECRYPT 2013*, pages 419–426.

Zwattendorfer, B., Stranacher, K., and Tauber, A. (2013). Towards a Federated Identity as a Service Model. In *Egovis 2013*, pages 43–57.

Interoperability for Web Services based Smart Home Control Systems

Hannu Järvinen and Petri Vuorimaa
Department of Media Technology, Aalto University School of Science, Espoo, Finland
{*hannu.jarvinen, petri.vuorimaa*}*@aalto.fi*

Keywords: Ambient Intelligence, Building Automation, oBIX, REST, Rule Engine, Smart Home, Web Intelligence, Web of Things, Web Services.

Abstract: One of the problems in smart home systems today is the lack of interoperability on different levels. While applying closed, non-standard, and complex protocols can cause the problem on a lower level, the architecture and design of a common building control system can cause it on a higher level. We present a solution for enabling the interoperability on the higher level in building automation systems with XML based rules and a Web API. To ensure the interoperability, we define requirements for Web services based building automation control systems. A standard building automation guideline, oBIX, is used to provide interoperability on the low level, and adopted for the rule management and description on the high level. The resulting rule engine architecture and implementation are evaluated against the requirements. The solution provides interoperability using standard Web technologies and supports employing several control systems simultaneously.

1 INTRODUCTION

In the Ambient Intelligence (AmI) world, our living environments consist of distributed embedded devices cooperating with one another while the intelligence of the system resides hidden in the network. These environments comprise of different kind of technology required to properly function together towards common goals. To achieve true interoperability between these technologies, four layers of interoperability need to be addressed: protocol layer, logic layer, semantic layer, and intelligence layer. In this paper, we present a solution for enabling smart home interoperability on the logic layer, and for supporting it on the semantic layer.

Common approach in smart home research has been to connect devices to a central control system that functions as a home server. Such a system can offer remote access interface for the users to manage their devices. The system can also offer means for a user to program certain logic by defining rules between the device states. While these kind of systems can help people to better manage their homes, they share two common constraints. In many cases, they are using closed and non-standard protocols for the communication, which restricts interoperability with generic devices. They also handle their logic internally in the control system blocking third party solutions out and forcing users to only use what is in hands. Hence, the system can function well with the devices that are connected to it, but is unable to interact with other systems or consider a new type of logic when built in functionality is felt insufficient. These independent systems can easily manage their internal behavior but the inter-system communication and logic is still a challenge.

In this paper, we present a solution for enabling higher level interoperability in smart homes. After a short introduction to home automation, we first discuss related recent research on smart homes and rule based systems, and then, considering the domain specific needs, define the requirements for Web services based building automation control system with basic logic interoperability and support for semantic interoperability. Next, we describe our architecture, use case and implementation, and finally evaluate the results against the requirements. The main contributions of our work are the requirements for the Web services based building automation control systems, and the rule engine implementation with high-level oBIX XML abstraction for the rule definition and management through a RESTful Web API.

2 BACKGROUND

Intelligent buildings have a long history and first attempts to make our homes smarter were planned al-

ready in the 1940s (Mozer, 1998). After that, the industry and research have periodically become interested in the idea of smart homes, and more lately the concepts of ubiquitous computing, pervasive computing and physical browsing have hyped the discussion on ambient intelligence up again.

Recent advancements in communication technology have brought us with possibility to network our home and building automation systems. The Internet is considered ubiquitous in modern buildings and provides a natural way to connect these systems together and with the rest of the IT world, offering the basic network infrastructure on both local and global scales. While the technology has been ready for smart buildings for a long time, they have not yet become reality for the people. Main reason for that has been the long battle between the standards on both national and international levels (Felser and Sauter, 2002). Agreements are not done on neither level, and today we have a huge set of competing device protocols.

The diversity of the device protocols forced to develop alternative solutions for modern Building Automation Systems (BASs). To cope with numerous non-interoperable standards, developers implemented first protocol converters between different devices, and later integration platforms, which are able to communicate over a variety of different protocols, and thus function as central servers for the device communication. An integration platform is basically a combination of hardware and software that is able to convert messages from different formats to others and communicate over different physical layers using a number of protocols (Järvinen et al., 2011) (Valtchev and Frankov, 2002). It can offer a high level interface to support information exchange with enterprise applications. Usually this interface, and more and more often the backbone network of the integration platform, is based on the Internet Protocol (IP) (Maile et al., 2007). Indeed, IP based communication is already widely applied in home automation systems. This allows exploiting successful and well-known Web technologies and tools in this domain.

Integration platforms partly solve the interoperability problem bringing information and controls from the devices available to the BASs. They unify numerous different protocols and data representations under one format, hopefully an agreed standard. However, after the interconnectivity, more is needed to make the devices to actually interoperate with each other. In practice, the interoperability on this level is realized with some kind of logic that makes decisions and strategies to execute commands based on the available information. In this paper, we refer to any system dedicated on handling this kind of logic in a building as a *control system*. Such system perceives the relevant environment states through, for example, current device state values, device history data, weather information and forecasts, or alternation of electricity price depending on the time of the day. It can also use any other external data available or a prediction made based on any of the mentioned data. Internal logic of the control system then evaluates this information with some reasoning and can decide to take actions based on the results.

In literature, various types of control systems have been studied. In addition to simple rule definition, other approaches such as fuzzy rules (Rutishauser et al., 2005) and Artificial Intelligence (AI) (Mozer, 1998) based solutions are suggested. What is common in traditional approaches is that they concentrate only on one type of control and handle it in a closed module with a specific user interface to manage the functioning.

The success of Web services as a solution for the integration of distributed heterogeneous systems has capitalized the research of service-oriented communications in the last decade. Web services are nowadays a *de facto* standard used to solve interoperability problems in distributed computing, which has motivated us to support interoperability offering a Web API for the rule management. Thus, in addition to devices, also the logic is controllable through the Web services interface.

In this paper, we concentrate on rule based control systems. However, we suggest supporting a whole horizontal layer of control systems simultaneously, thus enabling the interoperability also between the control systems. The focus of our research is on XML based rules in a distributed environment. In the next section, we describe earlier research on that area.

3 RELATED WORK

Rule based control systems generally describe rules with events, conditions, and actions. Such systems are able to automatically perform actions based on the occurring events and evaluated conditions. The idea is to have central management of the application logic, instead of implementing the functioning in separate programs, and to describe the rules with high-level, declarative syntax (Papamarkos et al., 2003). The composed Event-Condition-Action (ECA) rule structure has its roots in active databases. Originally, ECA rules were developed in eighties as a solution for the inability of the database management systems to trigger actions based on the alarms. Later on, ECA rules have been adopted by various other domains includ-

ing XML, semantic web, and ubiquitous computing (Chakravarthy and Adaikkalavan, 2007).

Considering the event part particularly in XML documents, W3C has standardized Document Object Model (DOM) Event Module (Schepers et al., 2011) for that purpose. In addition to UI events, it defines mutation events which are triggered on any change to the structure of a document. Corresponding event handlers can thus be created in a standard manner to handle required actions when the DOM is altered.

Similarly for the condition and action part, the Extensible Stylesheet Language Transformations (XSLT) (Kay, 2007) provides a standard way for applying rules to XML documents. It can be used to create an XML document based on two source documents. The source documents are an input XML data file, and an XSLT rule file describing which elements and attributes of the input XML are included and how the output file is then constructed. The application area of XSLT is well defined and narrow, thus offering a common tool for the application development. Additionally, related to our solution, extensions such as XSLT 2.0 Extensions for Saxon[1] and EXPath HTTP Client[2] offer support for making HTTP requests from the XSLT.

For the semantic layer, W3C has a recommendation for Rule Interchange Format (RIF). It has been developed for the exchange of rules between the rule languages, such as the oBIX based rule language we propose in this article. RIF thus maps different rule languages together using a common syntax and semantics. Our approach for the interoperability here however focuses on the logic layer providing an open way for managing the rules through a Web API. Together with addition of semantic layer, rule engine would benefit from RIF, as it would be interoperable also with other RIF supported rule engines.

One common format for describing rules in XML format is the Rule Markup Language (RuleML). (Boley et al., 2001) It defines a general and flexible rule language and is specified by a group of international experts on the field. However, in this paper we wanted to concentrate on the smart home control systems, and to unify the markup and the interface for both the devices and rules.

A number of approaches for applying ECA rules on XML has been proposed. Typically, they consider XML as a data description format for the database, otherwise keeping a close relation to the traditional database systems. This means that the rules operate on XML documents but are themselves described in various non-XML formats (Bailey et al., 2002) (Papamarkos et al., 2003) (Bonifati et al., 2002), and that the operations are done with a local database.

An approach from (Bonifati et al., 2001) defines the rules in XML, but also operates on XML documents locally. As the target documents are local, the system does not offer interoperability for distributed systems as we propose. Another approach from (Bernauer et al., 2004) tackled the problem of deciding on which events to react. As every modification of the document triggers Document Object Model (DOM) event, it is difficult to decide when conditions needs to be evaluated. Thus, they proposed a solution for detecting composite events for XML. However, they concentrate on the object node relations inside an XML document, and on the DOM events. This binds the solution to the local and complete XML documents. An article from (Papamarkos et al., 2004) presents an ECA language for Resource Description Framework (RDF). The language is able to function in distributed environments, as it supports path expressions with requests to resource URIs. The language itself is described in non-XML format, and the rule management is not discussed.

A recent research from (Leong et al., 2009) concentrated on rule based interaction. They employed ECA rules to enable interoperation between different heterogeneous subsystems in a smart home environment. They used SOAP as a backbone network protocol and stored rules in table form on a home application server. Subsystems reported their state changes to the home application server, which then fetched the related rules from the database, evaluated them and informed the subsystems if necessary. They defined an SQL based Application Programming Interface (API) for the management of the rules. The system could handle interoperation between the subsystems, but due to an additional API for the rule base, management of the rules could only be done using a specific user agent, not by other system modules.

4 REQUIREMENTS

Our approach is to gain the requirements for the Web services based building automation control systems from the properties of an Intelligent Agent (IA) defined in the Multi-Agent System (MAS) domain. An intelligent agent is a piece of software that is situated, autonomous, reactive, proactive, flexible, robust, and social (Padgham and Winikoff, 2004). We recognize four main properties of an IA that match with the properties needed for the control logic of a building automation system. These properties are: *autonomous* control, *reactive* responding to the changes

[1] http://www.fgeorges.org/xslt/saxon-ext/

[2] http://www.expath.org/modules/http-client/

in the environment, *proactive* observation of the environment, and *social* interaction with other agents.

We have derived the requirements from these properties and divided them into four categories. We suggest that the control descriptions should be based on standard XML description format, system be able to operate in distributed environments, support active knowledge base updates, and offer a standard interface for the management.

4.1 Description Format

To support social interaction, and for the simplicity of the control description, there are two requirements. The logic provided by the system should be described based on a common language utilized on the Web based building systems, similarly as all the data in the system should be presented using one standard XML based format. An XML based format is preferred due to its popularity in the Web, flexibility, both human and machine-readable nature, and direct support with Semantic Web techniques.

The description of the control logic in XML offers a way to hide the complexity behind a high-level abstraction layer. If a standard XML based building system language is utilized, there is no need to learn and implement new language syntax and semantics. Using a standard language also facilitates social interaction within different system modules and with the third party solutions. In addition to the XML encoding, the control logic should be described in a simple fashion. Declarative definition style defines the logic without describing the control flow. It contributes to the simplicity by offering a natural and intuitive way for presenting the logic. (Lloyd, 1994)

- R1: high-level standard XML abstraction for the control logic description
- R2: declarative control logic definition

4.2 Interaction in Distributed Environments

By their very nature, Web services based BASs are distributed device environments. In such environment, control systems should offer functionality for creating interoperability between various devices that provide their data and controls on the Web. Social interaction of the control logic and the devices can realize this behavior. Thus, control systems need to be able to both access web resources that are referred in their logic with URIs, and similarly to send requests to URIs for controlling web resources based on the following actions. A control system thus keeps up its knowledge base by perceiving the environment through references to multiple distributed web resources, instead of referring to path expressions within one XML document. This active inter-device communication also enables the interoperability in the action part of the logic. If the environment meets a certain state that changes the output of the logic, firing actions can reactively make requests to web based device resources, or similarly command other web resources, such as control systems.

- R3: ability to perceive the environment with web resource requests
- R4: ability to command actions with web resource requests

4.3 Knowledge Base Updates

Distributed nature and real-time requirements of building control systems require some considerations to be made for the knowledge base of the control system. One such requirement is the support for getting informed, or proactive/automatic updates of the web resource states that the system is aware of. A control system thus needs to be able to periodically request web resources for updating the knowledge base containing the relevant current perception of the environment.

- R5: active knowledge base updates

4.4 System Management

For the general interoperability, the control system management should be supported through a standard Web services interface. In addition, direct support for URIs as an addressing method simplifies the integration with the Web, and partly enables the social interaction. It also ensures that the logic is manageable the same way than other objects in the Web of Things (WoT) (Guinard and Trifa, 2009). While the control system expects devices to expose their functionality for the interaction, it should apply the same guideline to itself and provide its services through a Web API.

These requirements ensure that the application logic implementing the interoperation between the devices and subsystems is accessible for all system modules. Hence, the control modules of the system are manageable using the same standard interface than all the other modules of the system. This also allows logic to access other logic in the same control system, or in another type of control system.

Following these requirements in various control systems allows different modules to build a uniform

system, where management of any module is accomplished following the same consistent principles. In practice, modules can exploit the functionalities offered by other modules, thus supporting interoperability on a higher level.

- R6: control system management through standard web interface
- R7: logic identifiable with URIs.

5 ARCHITECTURE

We have defined an architecture for smart home systems following the requirements from the previous section, and apply those across all the system modules. As depicted in the Figure 1, on the high level all modules of the system are using the same Web services interface for the communication between them. This allows flexible management of the modules to achieve a configuration of the system that the user desires. Opposed to more complex SOAP approach, we promote use of RESTful Web services to simplify the communication and the design, and to facilitate the development of the following implementations. It is notable, that the applied Web service technique has to be open and standard to achieve interoperability with third party solutions.

As the communication interfaces are unified into one standard RESTful Web API, it is easy for modules to communicate with each other. In our model, this is the only interface for the management of a certain module. This important design principle assures, that the management of the modules is open for the whole system. As seen in the Figure 2, this is a conceptually different approach than having, for example, the application logic managed inside a closed module that is configurable only by a specific user interface or application. Now, control logic can be easily created and modified using the Web API that the modules provide.

While usage of only one interface makes the system simple, it also sets specific requirements for it. The interface standard has to allow a flexible way for developers to define their own diverse data models for describing their data. It has to be flexible enough to also allow the management of the control systems and be able to describe their functionality sufficiently. The network includes various different types of modules, and there can be multiple instances of any module type functioning at the system simultaneously.

An integration platform brings the information from the home devices available to the system. It communicates using various underlying low-level device specific protocols and unifies their representations to a standard data format offering an interface

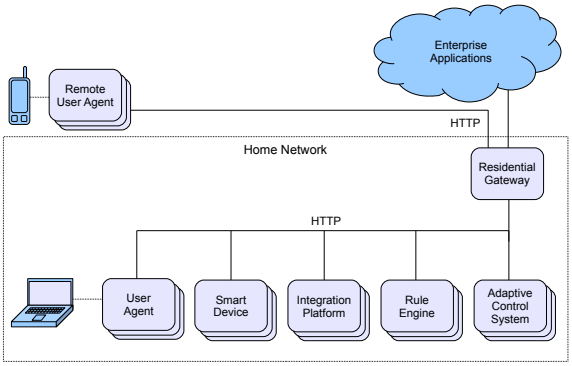


Figure 1: A smart home system model with distributed intelligence. A standard Web services based building automation guideline is applied for the communication.

for managing the devices that are connected to it. Smart devices instead are able to directly communicate using the standard backbone network protocol applied in the system. They do not need to be connected to any integration platform but if desired, they can do that as well. A smart device can also include application logic and it can, for example, directly communicate with other modules.

A rule engine is one type of control system. Using the standard RESTful Web API, other modules can create, modify and delete rules that run on the rule engine. Rules are a way to define interoperating logic to the system. Actions defined in an individual rule are executed if a certain logical condition is fulfilled. The rule engine then commands the related devices to take the appropriate actions. Rules are the most basic form of device independent interoperation in the smart home system. Similarly to a rule engine, adaptive control system is responsible for implementing interoperability to the system. The type of the logic, however, differs from the one in rule engines. The goal of the module is to learn from the behavior of the system and adapt the functioning to it. It can, for example, analyze the device history data and try to improve the efficiency and the user comfort. Because of this, the actual resulting behavior is not directly manageable but the related parameters are.

6 IMPLEMENTATION

The implementation focused mainly on the rule engine and the smart device but also included an integration platform and a couple of devices. As depicted in the Figure 3, the rule engine played a central part in the system. The following scenario was applied to test the implementation:

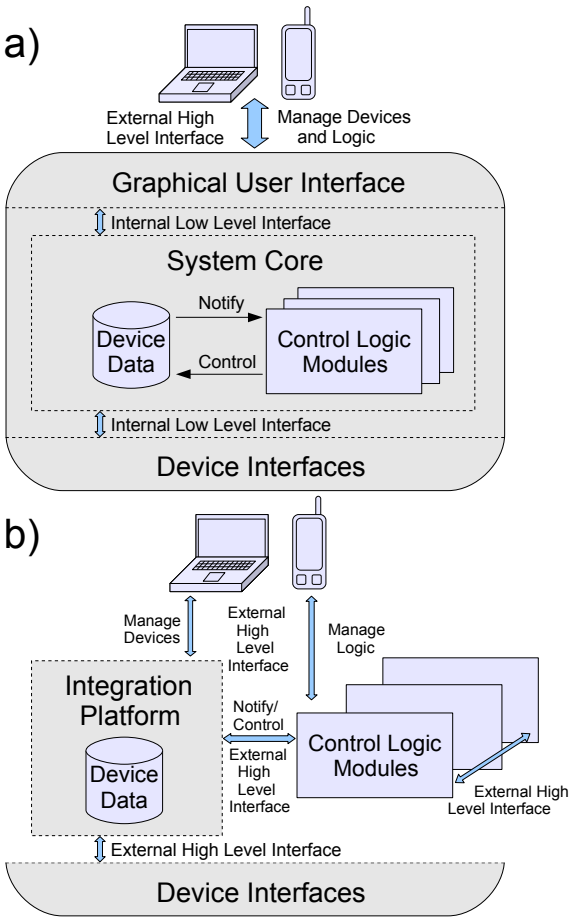


Figure 2: Communication between system modules in a system using a) Closed Control Logic; b) Open Control Logic.

Michael is watching television at the living room but wants to arrange some things at home at the same time. When he enters a room to pick up something, the light goes on in the room. When he leaves the room to put the thing in its correct place, the light switches off. However, when wandering around the flat he does not miss anything on the television, as always when he leaves the sofa, the program pauses. Similarly, when he sits down to the sofa again, the program continues.

For our implementation, we decided to use the open Building Information eXchange (oBIX) (Considine and Ehrlich, 2006) from the Organization for the Advancement of Structured Information Standards (OASIS). The data format is based on XML and defines, similarly to any common programming language, a small set of primitive data types for describing the data. It supports definition of contracts based on the primitives to allow developers to define data structures for their own devices and systems. The

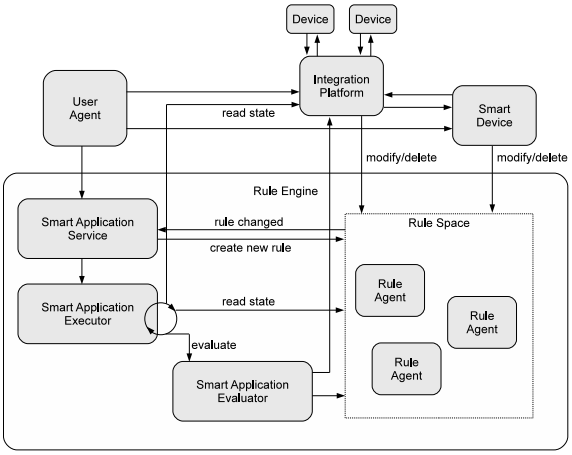


Figure 3: Communication flow of the implemented system.

extendibility suits well to our needs, as we want to describe multiple types of devices and control systems. The specification defines both SOAP and HTTP REST bindings. To simplify the development of the applications, we have selected to use the latter. An additional protocol layer is left out and the information can be directly read using an HTTP browser.

The environment had two rooms, both having three lights. For connecting devices to the integration platform, we decided to follow the same communication method than the rest of the system. Thus, the integration platform works as a standalone server and devices can connect to it through the Local Area Network (LAN). A camera based tracking system was used to get the location of the user. The tracking was implemented as a smart device and included additional logic. It used Linear Discriminant Analysis (LDA)[3] to predict if the user was about to enter the room or not. It was able to anticipate this behavior of the user and directly communicate with other system modules. The purpose of the rule engine was to allow interoperation in the system using simple logical rules, referred here as *smart applications*.

The idea of the prediction in the tracking system was to remove the general system delay so that the interaction with the environment would be pleasant. Rule agents provided backup functions for the light control, in case that the prediction would fail to recognize the behavior. In short, the overall system logic switched the lights on in the room which was entered and off in the room that was left. This happened exactly when the user entered the room without a delay. Additionally, the logic played and paused a TV program based on the user's presence on the sofa. The implemented system modules and especially the

[3]http://www.psychometrica.de/lda.html - Java implementation of the LDA.

different parts of the rule engine are described in more detail in the following subsections. Communication between these different system parts is emphasized. Detailed description of the user tests and measurements are presented in (Järvinen and Vuorimaa, 2012).

6.1 Rule Engine

An oBIX standard based rule engine was developed for the implementation. It is a further development result of our Java based open Facility Management Server (oFMS) (Järvinen et al., 2011). Thus, in addition to the rule management, the rule engine can simultaneously function as an integration platform. However, in this experiment we wanted to use a separate integration platform to follow the proposed architectural design.

The rule engine is not following the ECA structure as there is no separate event part in the rules. Its active component, the smart application executor takes care of regularly updating all the individual device states that are referred in any of the rule conditions. Thus, the conditions are evaluated only when needed. The actual triggering event is a change in any of the device states referred in the condition. The Algorithm 1 illustrates the behavior of the six rules that were used in the experimentation.

6.2 Smart Application Service

Smart application service offers an access point for external modules to create rules. Following the oBIX principles, it publishes information in the oBIX lobby object, access point to any oBIX based system (Figure 4). After the creation of a new smart application, all the needed references to it are found in the returned smart application oBIX object. This object provides functionality for management of the rule conditions and actions and the logical parameters affecting the

```
<obj href="obix:Lobby">
  <ref name="about" href="about/"
       is="obix:About"/>
  <ref name="smartApplicationService"
       href="smartApplicationService/"
       is="smartApplicationService"/>
</obj>

<obj href="smartApplicationService">
  <op name="make" href="make" in="obix:Nil"
      out="smartApplication"/>
</obj>
```

Figure 4: The smart application service is found in oBIX lobby. The service provides a way to create smart applications.

Algorithm 1: Rules used in the experimentation.

Rule 1: When someone enters or leaves this room, switch on and off the lights accordingly

if $rulestate.has.changed$ **then**
  if $rulestate = true$ **then**
    $lamp1 \leftarrow on$ **and** $lamp2 \leftarrow on$ **and** $lamp3 \leftarrow on$
  else
    $lamp1 \leftarrow off$ **and** $lamp2 \leftarrow off$ **and** $lamp3 \leftarrow off$
  end if
end if

Rule 2: When someone enters or leaves this room, switch on and off the lights accordingly

if $rulestate.has.changed$ **then**
  if $rulestate = true$ **then**
    $lamp4 \leftarrow on$ **and** $lamp5 \leftarrow on$ **and** $lamp6 \leftarrow on$
  else
    $lamp4 \leftarrow off$ **and** $lamp5 \leftarrow off$ **and** $lamp6 \leftarrow off$
  end if
end if

Rule 3: When the user moves to this room, notify the light batch rules and start observing when the user goes back to the other room

if $user.at.room.1$ **then**
  $rule1.rulestate \leftarrow on$ **and** $rule2.rulestate \leftarrow off$ **and**
  $rule4.active \leftarrow on$ **and** $rule4.statechangestolive \leftarrow 1$
end if

Rule 4: When the user moves to this room, notify the light batch rules and start observing when the user goes back to the other room

if $user.at.room.2$ **then**
  $rule1.rulestate \leftarrow off$ **and** $rule2.rulestate \leftarrow on$ **and**
  $rule3.active \leftarrow on$ **and** $rule3.statechangestolive \leftarrow 1$
end if

Rule 5: When the user sits on the sofa, turn on the TV and start observing when the user leaves the sofa

if $user.sitting.on.the.sofa$ **then**
  $tv \leftarrow on$ **and** $rule6.active \leftarrow on$ **and**
  $rule6.statechangestolive \leftarrow 1$
end if

Rule 6: When the user leaves the sofa, pause the program and start observing when the user returns to the sofa

if $user.not.sitting.on.the.sofa$ **then**
  $tv \leftarrow off$ **and** $rule5.active \leftarrow on$ **and**
  $rule5.statechangestolive \leftarrow 1$
end if

rule behavior. Smart application service is able to restore smart applications from the XML presentations stored in the database.

6.3 Smart Application Executor

Rule conditions can include references to a number of device and object states. The main task of the smart

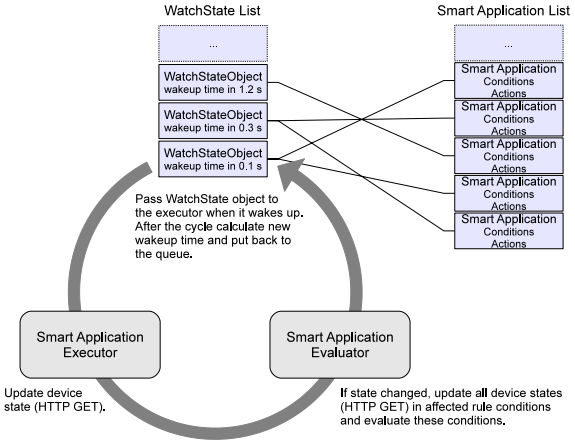


Figure 5: Smart application executor periodically updates the watchstate objects. If an object state has been changed, it creates a smart application evaluator to evaluate the conditions that refer to this watchstate.

application executor is to take care that these states are kept up to date. Based on the polling intervals defined in the rules, it periodically requests updates for the state values. When any of the states change, it creates a new smart application evaluator to evaluate the rules that include conditions that were affected (Figure 5).

6.4 Smart Application Evaluator

Smart application evaluator is spawned for every change in any of the device states that the smart application executor is updating. Executor passes a list of affected smart applications to the evaluator. Affected smart applications are all the rules that refer to the changed state in their conditions. Evaluator then goes through all the smart applications in this list evaluating their conditions. For the evaluation, it requests state values of the devices. If the return value of the evaluation differs from the current smart application state stored in its XML representation, corresponding rule actions are executed.

6.5 Rule Agent

In practice, rule agent describes a rule as a smart application (Figure 6). An individual rule includes rule properties defining its behavior, a conditions clause, and a list of actions. A conditions clause can include many condition clauses. All watch state rules in an individual condition clause have to be true for the rule condition to be fulfilled. Instead, for a conditions clause to be fulfilled, even one fulfilled condition clause is sufficient. An individual watch state object inside a condition defines a primitive condition for one device state (Figure 7). The resulting simple or-and mechanism can handle basic logic needed for common purposes. For more complex logic, rules can be chained together. This type of chaining is possible because rules are able to refer to the states of each other. Actions list defines the resulting actions that are executed if the state of the smart application is changed.

```
<obj is="smartApplication">
  <int name="stateChangesToLive"/>
  <bool name="active"/>
  <bool name="manualControl"/>
  <bool name="falseStateChange"/>
  <bool name="currentState"/>
  <reltime name="pollInterval"/>
  <op name="makeCondition" in="obix:Nil"
      out="condition"/>
  <op name="deleteCondition" in="obix:ref"
      out="obix:Nil"/>
  <op name="makeWatchState" in="obix:ref"
      out="watchState"/>
  <op name="deleteWatchState" in="obix:ref"
      out="obix:Nil"/>
  <list is="conditions" of="condition"/>
  <op name="makeAction" in="obix:Nil"
      out="action"/>
  <op name="deleteAction" in="obix:ref"
      out="obix:Nil"/>
  <list is="actions" of="action"/>
  <op name="delete" in="obix:Nil"
      out="obix:Nil"/>
</obj>
```

Figure 6: The smart application contract in oBIX. In practice, smart application instances also include href attributes, so that the individual elements can be referred. Thus, each rule and its sub elements have unique URIs.

```
<obj is="watchState">
  <str name="target"/>
  <enum name="relation"
      range="relationshipRange"/>
  <val name="ComparisonValue"/>
  <reltime name="pollInterval"/>
  <obj name="state">
    <bool name="active"/>
    <str name="message"/>
  </obj>
</obj>
```

Figure 7: An oBIX contract for a watch state object defining a simple condition which refers to only one device state. Instances of the contract also include href attributes for each of the elements.

6.6 Integration Platform

We used an integration platform provided by C oBIX Tools (CoT)[4] from Andrey Litvinov. It handles all the

[4]C oBIX Tools is published as an open source project at http://code.google.com/p/c-obix-tools/

basic oBIX features and additionally offers a more efficient technique for polling than the original oBIX specification. The integration platform was used by the lighting system oBIX client, virtual television oBIX client, and, of course, the rule engine and the smart device for the interoperation.

6.7 Smart Device

The tracking system oBIX client was implemented as a smart device. It was able to predict the user entering into a room and managed to compensate the common system delay of approximately one second in that behavior. It used LDA for the prediction when the user was close to the door. LDA considered walking direction and the turning angle as factors to define the behavior. Smart device did not command the lights directly but used the batch rules (Rules 1 and 2) on the rule engine to easily command multiple lights in one room with one request.

7 EVALUATION

The results are evaluated against the requirements for Web services based BA control systems defined in Section 4.

7.1 Description Format

The control system logic is implemented as rules. The system is based on the oBIX standard, which defines an XML-based data description format. The use of oBIX XML data model for the rule description fulfills the requirement R1 per se. This language is used to define rules in a declarative manner by describing them with condition clauses, watch states, actions, and general rule parameters, fulfilling the requirement R2. For further interoperability to exchange rules with other systems, RuleML could be used as a rule description format, and the rule management interface could handle the translation into oBIX for the Web API.

7.2 System Management

In the XML description, a rule element and its sub elements all have unique URIs. Thus, they are manageable through the HTTP oBIX interface. This direct support for URIs as an addressing method simplifies the rule management, provides an access method for the control systems to perceive and control the environment, and facilitates the integration with the rest of the Web. These fulfill the requirements R6 and R7.

7.3 Active Interaction in Distributed Environments

The implemented smart application executor regularly requests updates for the device states that the rules depend on, fulfilling the requirement R5. Additionally, the rule engine contains a smart application evaluator, which evaluates the rule conditions and commands the actions if the conditions are fulfilled. Both the executor during the regular update cycle and the evaluator during the rule evaluation and action commanding, are able to make HTTP requests to Web resources. These fulfill the requirements R3 and R4.

8 DISCUSSION

According to Zelkha, et al. (Zelkha et al., 1998), the ambient intelligence paradigm is characterized by the systems that are:

- Embedded;
- Personalized;
- Adaptive;
- Anticipatory.

The system presented in this paper can be characterized with these features. The integration platform provides a solution for interconnecting the devices embedded in the environment. Functioning of the devices can then be managed and the system personalized through standard BA interface, and automatic behavior be programmed with the rule engine. Utilizing a standard interface enables usage of multiple control systems at the same time, thus allowing modules for adaptive and context aware control to be added to the system. Similarly, the system demonstrates the possibilities for anticipatory behavior with a simple prediction module for lighting control.

We believe that the defined system requirements offer a basis for developing interoperable smart home systems. From the architectural point of view, the main advancement is the support for multiple control systems at the same time. While it offers a possibility for modular system design, it also introduces new challenges for the management of the system functioning as a whole. Cooperative functioning of this kind of distributed applications have been studied in the MAS domain (Weiss, 1999) (Wilmott et al., 2001). Thus, applying advanced MAS techniques and principles for the flexibility, robustness, and conceptual reasoning of the control system agents is one of our future research lines. In contrast to similar existing systems, our solution is based on Web technolo-

gies and provides a Web API for the rule management, which allows user agents to create and remove rules, but also to modify the content of existing rules in the rule engine. These with an XML based description format are compatible with Semantic Web techniques and allow us to add a new semantic layer on top of the system in the future, and integrate MASs with the smart home systems to further build intelligence with semantic reasoning.

9 CONCLUSION

We presented a logic layer interoperability solution for distributed smart home systems. The system was based on requirements that ensure the support for features that contribute to the interoperability on the logic layer and enable it on the semantic layer. The implementation considered a rule engine able to observe and control devices through standard Web services interface. Similarly, the rule engine offered a Web services interface for the rule management following the same building automation guideline. The implemented system was functional, fulfilled the requirements, and proved the applicability of the approach. However, the system does not offer all the needed functionalities at the moment. Proper methods are needed for security, authentication, authorization as well as managing conflicts in the logic. As the system is using RESTful Web services, it can partly exploit same well-known techniques that are used in the traditional Web applications to overcome some of the challenges.

We believe that Web services based control systems can offer solutions to gain better interoperability in smart home environments. The interoperability in the system is not achieved by offering a single manner to control the system. It is a joint effort of multiple control systems offering their services openly for managing the logic. Individual control systems can concentrate on their main functions and exploit services from each other, shaping the smart home to a fully functional ecosystem where every service does its own part and interacts with the others.

ACKNOWLEDGEMENTS

Part of the work was done as part of TIVIT Devices and Interoperability Ecosystem (DIEM) project Building Automation work package and supported by The Doctoral Programme in the Built Environment (RYM-TO) funded through the Academy of Finland and the Ministry of Education and Culture.

REFERENCES

Bailey, J., Poulovassilis, A., and Wood, P. (2002). An event-condition-action language for xml. In *Proceedings of the 11th international conference on World Wide Web*, pages 486–495. ACM.

Bernauer, M., Kappel, G., and Kramler, G. (2004). Composite events for xml. In *Proceedings of the 13th international conference on World Wide Web*, pages 175–183. ACM.

Boley, H., Tabet, S., and Wagner, G. (2001). Design rationale for ruleml: A markup language for semantic web rules. In *SWWS*, volume 1, pages 381–401.

Bonifati, A., Braga, D., Campi, A., and Ceri, S. (2002). Active xquery. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 403–412. IEEE.

Bonifati, A., Ceri, S., and Paraboschi, S. (2001). Active rules for xml: A new paradigm for e-services. *The VLDB Journal*, 10(1):39–47.

Chakravarthy, S. and Adaikkalavan, R. (2007). Ubiquitous nature of event-driven approaches: A retrospective view.

Considine, T. and Ehrlich, P. (2006). Open building information exchange (oBIX).

Felser, M. and Sauter, T. (2002). The fieldbus war: History or short break between battles? In *IEEE WFCS*, pages 73–80.

Guinard, D. and Trifa, V. (2009). Towards the web of things: Web mashups for embedded devices. In *Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web (MEM 2009), in proceedings of WWW (International World Wide Web Conferences), Madrid, Spain*.

Järvinen, H., Litvinov, A., and Vuorimaa, P. (2011). Integration platform for home and building automation systems. In *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pages 292–296. IEEE.

Järvinen, H. and Vuorimaa, P. (2012). Anticipatory lighting in smart building. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, pages 390–394. IEEE.

Kay, M. (2007). Xsl transformations (xslt) version 2.0. *W3C Recommendation*, 23.

Leong, C., Ramli, A., and Perumal, T. (2009). A rule-based framework for heterogeneous subsystems management in smart home environment. *Consumer Electronics, IEEE Transactions on*, 55(3):1208–1213.

Lloyd, J. (1994). Practical advantages of declarative programming. In *Joint Conference on Declarative Programming, GULP-PRODE*, volume 94, page 94.

Maile, T., Fischer, M., and Huijbregts, R. (2007). The vision of integrated IP-based building systems. *Journal of Corporate Real Estate*, 9(2):125–137.

Mozer, M. (1998). The neural network house: An environment hat adapts to its inhabitants. In *Proc. AAAI Spring Symp. Intelligent Environments*.

Padgham, L. and Winikoff, M. (2004). *Developing intelligent agent systems: a practical guide*, volume 1. Wiley.

Papamarkos, G., Poulovassilis, A., and Wood, P. (2003). Event-condition-action rule languages for the semantic web. In *Workshop on Semantic Web and Databases*, page 2003. Citeseer.

Papamarkos, G., Poulovassilis, A., and Wood, P. (2004). Rdftl: An event-condition-action language for rdf. In *Proc. of the 3rd International Workshop on Web Dynamics*.

Rutishauser, U., Joller, J., and Douglas, R. (2005). Control and learning of ambience by an intelligent building. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(1):121–132.

Schepers, D., Rossi, J., Höhrmann, B., Le Hégaret, P., and Pixley, T. (2011). Document object model (dom) level 3 events specification. *W3C Working Draft*.

Valtchev, D. and Frankov, I. (2002). Service gateway architecture for a smart home. *Communications Magazine, IEEE*, 40(4):126–132.

Weiss, G. (1999). *Multiagent systems: a modern approach to distributed artificial intelligence*. The MIT press.

Wilmott, S., Dale, J., Burg, B., Charlton, P., and O'Brien, P. (2001). Agentcities: A Worldwide Open Agent Network. *AgentLink News*, Issue 8.

Zelkha, E., Epstein, B., Birrell, S., and Dodsworth, C. (1998). From devices to ambient intelligence. In *Digital living room conference*.

An Integrated Approach for Designing and Validating REST Web Service Compositions

Irum Rauf, Faezeh Siavashi, Dragos Truscan and Ivan Porres
Åbo Akademi University, Dept. of Information Technologies, Turku, Finland
{*irum.rauf, faezeh.siavashi, dragos.truscan, ivan.porres*}*@abo.fi*

Keywords: REST, Web Service Composition, Model-based Testing, UPPAAL, TRON.

Abstract: We present an integrated approach to design and validate RESTful composite web services. We use the Unified Modeling Language (UML) to specify the requirements, behavior and published resources of each web service. In our approach, a service can invoke other services and exhibit complex and timed behavior while still complying with the REST architectural style. We show how to transform service specifications into UPPAAL timed automata for verification and test generation. The service requirements are propagated to the UPPAAL timed automata during the transformation. Their reachability is verified in UPPAAL and they are used for computing coverage level during test generation. We validate our approach with a case study of a holiday booking web service.

1 INTRODUCTION

Web services have machine-readable interfaces that automate the task of communicating information between machines and reduce time and human efforts. They are being increasingly used in the industry in order to automate different tasks and offer services to a bigger audience. REST architectural style aims at producing scalable and extensible web services using technologies that play well with the existing tools and infrastructure of the internet. This has encouraged notable enterprizes to use REST web services to meet their needs. REST interfaces offer a CRUD interface (create, retrieve, update and delete) to its users via a set of standard HTTP methods. They offer stateless behavior that facilitates scalability and requires that no hidden session or state information be carried between method calls.

Different web services published over the internet can be composed to form a composite web service such that the composed web service fulfills new service goals using the functionality of partner web services. Automated systems, for example hotel reservation systems, are often built as stateful composite services that require a certain sequence of method invocations that must be followed in order to fulfill service goals. Designing and developing such stateful composite services with REST features is not a trivial task and requires rigorous approaches that are capable of creating reliable web services.

Thus, with the use of web services in businesses and critical applications, there is an increasing need for a) design approaches to develop web service compositions that support complex scenarios and timed behavior while complying with the REST architectural style and b) validation approaches to effectively and efficiently detect faults in specifications and implementations of such services.

In this article, we present a design and validation approach that facilitates the service developer to create reliable, timed and stateful composite REST web services. As a first contribution, we introduce an approach in which we use the Unified Modeling Language (UML) (UML, 2009) to model our service specifications via an extension of our previous work in (Porres and Rauf, 2011). We use UML since it has emerged as a standard modeling language at industrial level (Budgen et al., 2011) and has sophisticated tools due to large user base. This can make adoption of the approach easier in the industry.

The service design models and their implementations should be validated for their correct behavior in order to build trust on the service functionality. We have used the model checking approach for this purpose. Model checking is a way to exhaustively and automatically check if a finite-state model of a program satisfies its specifications (Clarke et al., 1994). The UML-based service design models represent the system graphically and are comprehensible for a human user. In order to make the models amenable

for model checking we suggest, as a second contribution, a set of reversible mechanized steps for translating UML-based service specifications into UPPAAL timed automata (UPTA) (Larsen et al., 2009). UPPAAL is a commonly used model checking tool for verifying real time systems through modeling and simulation (Larsen et al., 1997). We verify basic properties of our design models such as reachability, liveness and safety using the UPPAAL model-checker tool. UPTA are updated (if needed) based on the verification results and transformed back to UML. From the UML models, a skeleton of the composite service is generated automatically in the Django web development framework (Holovaty and Kaplan-Moss, 2009) using our partial code generation tool.

In our approach, a service can invoke other services and exhibit complex and timed behavior while still complying with the REST architectural style. We need to check if the service implementation is functioning correctly along with partner services and if the service goals and timed constraints are being fulfilled. Thus, as a third contribution, we show how we validate the implementation of the RESTful web service composition with a model-based testing (MBT) approach using the UPPAAL TRON tool (Larsen et al., 2009). By using MBT, test cases can be automatically generated with an increased probability of test coverage and with an ease of test case maintenance.

Requirements traceability is an important component of our integrated approach. The requirements of the composition are included in the UML specifications and then propagated to UPTA specifications. They are used for both verifying the reachability of those model elements implementing them and for reasoning about their coverage after the tests are executed. Upon detecting failures, traced requirements can be used to localize the errors either in the models or in the specifications.

We exemplify and validate our integrated approach with a relatively complex case study of a holiday booking composite REST web service from industrial context. The case study shows how stateful and timed web services offering complex scenarios and involving other web services can be constructed using our approach.

The paper is organized as follows: Section II gives an overview of our approach, while tool support is discussed in section III. The case study is presented in section IV and the evaluation of the approach is presented in section V. The related work is discussed in section VI and the section VII concludes the paper.

2 OUR APPROACH

An overview of our integrated design and validation approach is given in Figure 1. The left side of the figure shows our previous work, whereas the right side shows the contribution of this paper and how it is connected to the previous work.

In our previous work (Figure 1–left), we designed behavioral interfaces for web services that were RESTful by construction (Porres and Rauf, 2011). These design models were implemented in the Django web framework using our partial code generation tool (Rauf and Porres, 2011) which generated code skeletons with pre- and post-conditions for every service method. The design models were also analyzed for their consistency (Rauf et al., 2012).

In the current work, we focus on designing, verifying and testing composite REST web service (Figure 1–right) as follows:

Design: First of all we extend our design approach to create composite REST web services with UML. Our approach takes as input the behavioral interface specifications of the partner REST web services and the business requirements. With these inputs, we construct models for composite web service using our design approach.

Verification: We then provide verification of the design models by reasoning on the basic properties of models like deadlock, liveness, reachability and safety with the UPPAAL model-checker. To achieve this, we transform the service design models to UPTA, which are simulated and verified. Based on verification results, the UML-based service design models are updated.

Transformation: Transformation step generates two types of automata from service design models. One of the types corresponds to service design models and the other type represents the environment model. The environment model simulates the behavior of service user to invoke the interface service methods in order to facilitate test generation.

Code Generation: The code skeletons are generated from UML service design models to Python-based Django web framework using our code generation tool (Rauf and Porres, 2011) that are completed manually by the developer.

Testing: For model-based testing of the service implementation, we have used online conformance testing tool UPPAL-TRON that validates the service implementation against its UPTA specification models at runtime.

Evaluation: In the end, we have evaluated our validation approach for its efficiency using mutation testing and benchmarked.

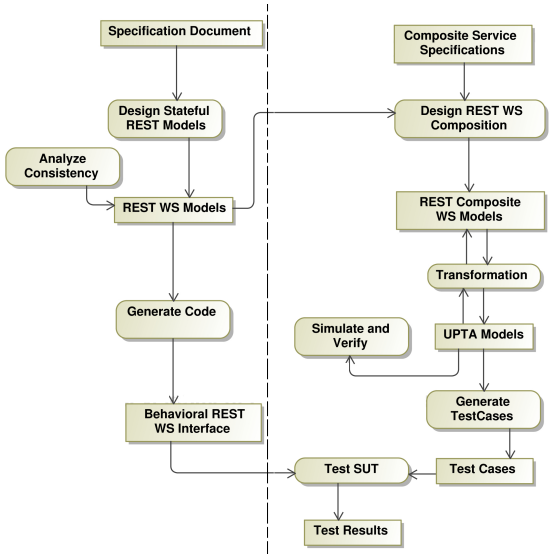


Figure 1: Activity Diagram of Design and Validation Approach for REST CWS.

2.1 REST Composition Models

We require that a composite REST web service interface should exhibit the REST interface features, i.e. addressability, connectivity, statelessness and uniform interface. We have modeled our composite REST web service interfaces with a *resource model*, a *behavioral model* and a *domain model* that exhibit these features.

This work extends our previous work on creating behavioral interface specifications of individual REST web services (Porres and Rauf, 2011). For the details of the design approach, readers are referred to (Porres and Rauf, 2011).

The concept of resource is central to the structure of REST web service. It represents a piece of information (Richardson and Ruby, 2008). We respresent the static structure of REST web service with resource model which is modeled with a UML class diagram. Each class represents a *resource definition*. We have used the term *resource definition* to define a resource entity such that its instances are called resources. This is analogue to the relationship between a *class* and its *objects* in the object-oriented paradigm.

The direction of the association between *resource definitions* gives the navigability direction between them while their role names give the relative URI of resources (addressability). The *collection resource definiton* without the incoming transitions is termed as *root* such that every *resource definition* in the resource model should be reachable via *root* and the graph formed should be connected (connectivity).

A behavioral model represents the dynamic structure of the service and it is modeled by a UML State Machine (SM). Each state represents the service state and the trigger methods of transitions are restricted to the side-effect methods of HTTP, i.e., PUT, POST and DELETE (uniform interface). The statelessness feature of the REST interface is preserved while building stateful REST web service by defining state invariants as boolean expression of states of different resources. The state of a resource is given by its representation retrieved by invoking a GET on it. We are thus able to define service states as predicates over the resources without maintaining any hidden session or state information (statelessness). The state invariants in the SM are written as Object Constrain Language (OCL) expressions. OCL is commonly used to define constraints in UML models, including state invariants (Birgit Demuth, 2009).

For modeling a service composition, the models are required to represent method invocations on the partner services. The service invocations to partner services are modeled as effects on the transitions. The composite web service requirements, inferred from the specification document, are added as UML *comments* on the transitions that satisfy them. These requirements should be met by the implementation of the service in order to fulfill the service goals.

The *domain model* of the composite service is represented with a class diagram. It represents interfaces between the composite service and its partner services. The required and provided interface methods between the composite and its partner services are modeled with required and provided interfaces in the domain model, respectively.

2.2 Verification

Model verification is a process of determining whether the models are designed correctly and represent the developer's conceptual descriptions and specifications. Model checking is one of the ways to exhaustively check the models automatically. The service design models of composite REST web service should be verified for their basic properties in order to build confidence of the service designer on the models before implementing them. This allows one to eliminate design errors that can be expensive to detect and correct at later stage of the development cycle.

UPPAAL model-checker is used for modeling, simulation and verification of real-time systems (Larsen et al., 1997). It consists of set of timed automata (TA), clocks, channels that synchronize the systems (automata), variables and additional elements. A real-time system is modeled as a closed network of TA. Each automaton in the network is speci-

fied via a *template*, which can be instantiated as *process*. A template in UPTA is composed of *locations*, *edges*, *clocks* and *variables* representing all properties of the system. Synchronization between different processes can be provided using *channels*. Two edges in different automata can synchronize if one is emitting (denoted as *channel_name!*) and the other is receiving (denoted as *channel_name?*) on the same channel. *Guards* are the conditions that enable a transition when they are satisfied. Similarly, the conditions associated to locations, called *invariant*, specify that the system can stay in the location if and only if the invariant is satisfiable.

The query language used in UPPAAL is a simplified version of TCTL (Alur et al., 1990) that consists of state formulae and path formulae. State formulae (φ) is an expression that describes an individual state, while path formulae can be classified into reachability, safety and liveness properties. Deadlock is expressed using state formulae. The syntax of TCTL path formulae that are used in UPPAAL is defined as follows:

- A $\Box$ φ - for all paths, the property φ is always satisfiable.
- A $\Diamond$ φ- for all paths, the property φ is eventually satisfiable.
- E $\Box$ φ - there is at least a path in the automata such that property φ is always satisfiable.
- E$\Diamond$ φ - there is at least a path in the automata such that property φ is eventually satisfiable.
- $\varphi \rightsquigarrow \phi$ - when φ holds, ϕ must hold.

If there is a location in the model that has no outgoing transition, then the model is said to be in a deadlock. Reachability properties validate the basic behavior of the model by checking whether a certain property is possible in the model with the given paths. The safety property checks that something bad will never happen and the liveness property is verified to determine that something will eventually happen.

However, before using UPPAAL model-checker to verify these properties we need to give our service design models in UML formal foundations that are understandable by the verification tool. This has to be done in an automated manner to avoid extra efforts from the service developer. In section 3, we present our tool support for this and explain in detail the transformation from UML to UPTA.

2.3 Model-Based Test Generation

Model-based testing (MBT) is a method that provides an abstract model of a system under test (SUT) and preforms automatic test case generation based on the specifications of the SUT (Utting and Legeard, 2007). In MBT, modeling the environment of a system is important since the environment generates test cases from whole or some parts of the model to satisfy the test criteria. Environment models help in automation of testing in three ways: the automation of test case generation from a simulated environment, the selection of test cases, and the evaluation of their test results. Our UML to UPTA transformation tool generates both the SM of SUT and the environment model.

We provide automatic test generation using UPPAAL TRON, which is an extension of UPPAAL for online model-based black-box conformance testing (Larsen et al., 2009). During test generation, the environment model randomly selects test cases and communicates to the test adapter.

A test adapter is used by UPPAAL TRON to expose the observable I/O communication between the test environment model and the SUT model, as shown in Figure 2. Our adapter implements the communication with the SUT by converting abstract test inputs into HTTP request messages and HTTP response messages into abstract test outputs. The TRON tool generates tests via symbolic execution of the specifications using randomized choice of inputs. Based on the timed sequence of input actions from the simulation, the adapter preforms input actions to Implementation Under Test (IUT) and waits for the response. Output from IUT is monitored and generated as output actions for the simulation. The conformance testing is achieved by comparing outputs of IUT to the behavior of the simulation.

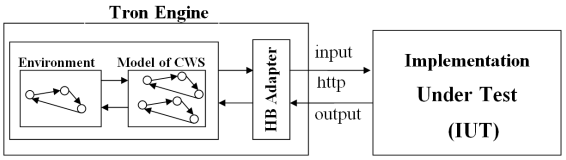


Figure 2: UPPAAL TRON test setup.

2.4 Requirements Traceability

Service requirements can be inferred from the specification document and they serve as service goals. A service should be checked for its service goals in order to validate that the service does what it is required to do. By catering to the service requirements at the design phase and propagating them to the validation stage, we provide a mechanism by which a service requirement can be validated for its goals and the unfulfilled requirements can be traced back to the design phase to find faults in the design. Service requirements are generally domain-specific since they are inferred from the specifications. We infer functional and temporal requirements from the specification document into a table and number them. These

requirements are attached to the SM as *comments* on the transitions and are propagated to UPTA such that the links between requirements and the model elements are preserved. These requirements are included in all the models and traced throughout the process, i.e., at UML, UPTA and test level, respectively.

The requirements are formulated as reachability properties in UPTA with the purpose of verifying them during simulation. Each requirement label is translated into a boolean variable (initialized to False) and attached to the corresponding edge in the UPTA. This mapping is explained in more detail in the Section 3 on the UML to UPTA transformation.

We require that our testing approach must validate that these requirements are met by IUT, in order to build confidence of the developer that the system is doing what it is required to do. Thus, their coverage level is monitored during test generation and execution. Once the test report is available, we can check which requirements have been validated and which have failed.

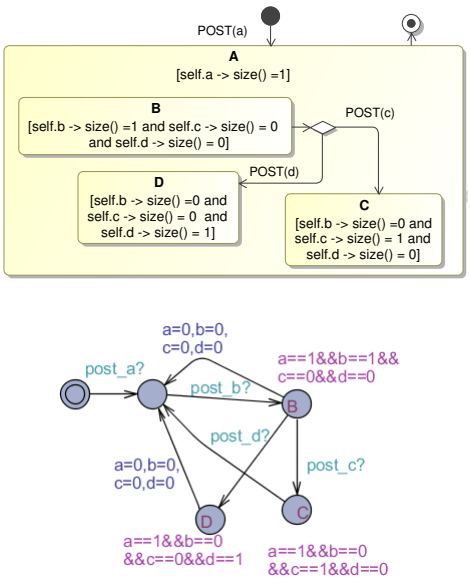


Figure 3: (Left) Composite State in UML State Machine. (Right) Flattened locations in timed automaton.

3 TOOL SUPPORT

Modeling in UML. The design models are modeled using MagicDraw (Mag, 2013). Static validation of models is done via OCL using the validation engine of Magic Draw. We rely on predefined validation suites for UML contained in MagicDraw for the basic validation of the model. These validation suites contain rules that check that the designed UML model conforms to UML meta-model specifications and prevent the developer from doing basic modeling mistakes.

Code Generation. The code-skeleton of the updated service design models of REST composite web service can be generated using our tool presented in (Rauf and Porres, 2011). The tool generates code skeleton for design models in Django that is a high level Python web framework (Holovaty and Kaplan-Moss, 2009). The generated code also has behavioral information such that the pre and post conditions for each method are included and the developer just has to write the implementation of the operations.

UML→UPTA Transformation. The transformation from UML design models to UPTA is an extension of our approach presented in (Nobakht and Truscan, 2013). The extension of transformation generates several artifacts: UPTA model, test environment model and a skeleton for test adapter depicting the testable interfaces of the composite web service.

Resource Model. In UPTA the resource model is represented as a template. The *resource definitions* in the resource model are specified as variables with 1 or 0 value, specifying if a resource exists or not, respectively. The attributes of *resource definitions* are inspected and for each integer attribute, an integer variable is declared in the UPPAAL model. Similarly, the boolean attributes are declared as integer arrays of 0 and 1.

Domain Model. The domain model shows set of operations offered and required by the composite web service and its partner web services. The corresponding communication between templates in UPPAAL is represented by channel synchronizations. Two edges in different automata in UPPAAL can synchronize if one is emitting and the other is receiving on the same channel. The operations in an interface are thus translated into a binary synchronization channel in UPPAAL. The template of the service that realizes the interfaces acts as the receiving automaton and the sending automaton is specified by the template of the service that uses the interface.

Behavioral Model. The SM of the REST web service is encoded to TA that are represented by templates, which are instantiated as processes. Figure 3 shows an example of transformation from the SM to TA.

States. A state is mapped to a location in UPTA, and a state invariant is mapped to corresponding location invariant. The subclauses of the state invariant are translated to variables corresponding to the respective resource definition. For example, in Figure 3, $self.a->size() = 1$ is translated as $a = 1$ and $self.b->size() = 0$ as $b = 0$. The initial state corresponds to the initial location. The final states are translated by having an edge from the corresponding

location to initial location and updating all the variables to their initial values, as shown in Figure 3. The choice state in the SM is replaced by two edges in the TA model that are originating from the same source location to different target locations.

State Hierarchy. The SM may contain composite states for better representation of specifications. UPTA, however, does not support the notion of location hierarchy. We flatten the composite states into several simple states by including the state invariant of super states in the contained states that are then mapped to the respective locations in UPTA. For example, in Figure 3, the top figure contains a SM with a composite state that is flattened to UPTA model shown at the bottom. States B, C and D in the SM correspond to the locations B, C and D of UPTA, respectively. Note that all the locations contain the state invariant of superstate A in the SM.

Transitions. A transition in the SM is mapped to an edge in UPTA and guards on the transition are mapped to guards on the corresponding edge in UPTA. In Figure 4, we shows how the transitions in the SM (top) are translated to UPTA (bottom right). The locations $L1$ and $L5$ correspond to states $S1$ and $S2$ of SM, respectively, and locations $L0$, $L2$, $L3$, and $L4$ are the extra locations created during the transformation process as explained below. The state invariants are translated to location invariants and represented as boolean functions for the purpose of diagram clarity. The transition between states S1 and S2 is triggered by POST(b) after 10 minutes as specified in the guard. In UPTA, this is represented as guard over the clock variable cl.

Trigger Methods. The trigger methods from the SM are translated in to receiving channels in UPTA. This receiving channel is in sync either with the automaton of the partner service or with the environment model.

Time Events. The time events in behavioral diagram are replaced by clocks in UPTA. The clock is reset in the incoming edge to the location (L1) and is also included in the location invariant. Thus, the guard *after(10m)* is translated to $cl > 10$ on the corresponding UPTA edge.

Effects. The effect on the transition, i.e., $POST(c)$ shows invocation to the partner service. The communication between two web services is established by using a unique channel synchronization. For instance, emitting a request from a web service to the other one can be replaced by synchronizing a channel in an UPPAAL process, and the other process is the receiver of the synchronization. The effect of the transition that invokes a remote service is represented with two edges and an urgent location (marked with U in the circle) in between, i.e., edges *e2* and *e3* and *urgent* location *L3*. An urgent location in UPTA does not allow any delays (Larsen et al., 1997). Thus, the first edge (e1) is synchronized with the environment model and the second edge (e2) synchronizes with the partner automaton. The third edge (e3) is synchronized to receive acknowledgment response from the partner (as we model asynchronous service) and the sending channel on the fourth edge (e4) is synchronized with the environment to indicate the completion of transition.

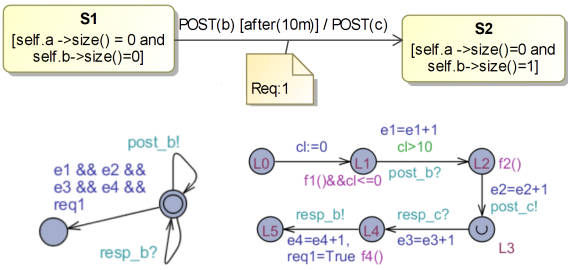


Figure 4: Example of SM (top) Corresponding Environment Model (bottom left) and Flattened TA (bottom right).

Requirements. The requirement on the transitions are translated into a boolean variable (initialized to *False*) and attached to the corresponding edge which updates it to *True*. This is shown in Figure 4 with *Req1= True* on edge e4. This implies that whenever this edge would be traversed, this requirement will be met. This can be formulated as reachability properties to attain requirement coverage and tracked during test generation and execution.

Environment Model. The environment model in UPTA has sending channels that are received by the composite web service automaton as inputs to trigger the process. This is similar to interface method calls in the SM. All the interface methods of the service specified in the state machine are mapped to the sending channels in the environment model and the response of successful transition is received from the composite web service via receiving channels. This is also shown in Figure 4: the environment model initiates the automaton (bottom right) by sending channel *post_b*! and the process completes when the channel *resp_b*? is received.

A Python script is currently used to create the environment model, from a given UPTA model by analyzing the channels observable from the environment. The original idea has been discussed in (Hessel et al., 2008). This will be merged in the final version of the UML→UPTA transformation script.

Test Coverage Information. In order to enable rigorous test coverage in UPPAAL TRON, a second Python script (discussed in more detail in (Koskinen et al., 2013)) is used to automatically add *counter*

variables for each edge of a given automaton in a UPTA model and a corresponding update of the given variable on the corresponding edge. Whenever the edge is visited during the simulation or execution, the variable is incremented, allowing thus to track which edges have been visited and how many times. This enables one to track coverage level wrt. e.g., edge coverage or edge pair coverage. This script will also be integrated in the final version of the UML→UPTA transformation script.

4 CASE STUDY

Our case study is a Holiday Booking (HB) composite REST web service that is built on inspiration from the *housetrip.com* service, with the purpose of having a case study similar in complexity to real services. This service is a holiday rental online booking site, where one can search and book an apartment in the destination country.

The user of the service searches for a room in a hotel from the list of available hotels at HB before travel. He books the room (if it is available) and that booking is reserved by HB with the hotel for 24 hours. The user must pay for the booking within 24 hours. If the user does not pay within this time then the booking is canceled. If the booking is paid, then the HB service invokes a credit card verification service and waits for the payment confirmation. When the payment is confirmed, HB invokes the hotel service to confirm the booking of the room. If the hotel does not respond within 1 day or it does not confirm at all, the booking is canceled and the user is refunded. If the hotel service confirms, then a booking is made with the hotel. The payment is not released to the hotel until the user checks in. When the user checks in, HB releases the money to the hotel and the booking is marked by the hotel as paid. Due to space limitation, we only show some of the models and information here. The detailed case study is available at (Rauf et al., 2013)

Design Models. The design of HB composite REST web service is modeled with resource, behavioral and domain models. The state machine of HB composite service is shown in Figure 5.

Requirements Traceability. We have inferred functional and temporal requirements from specification document for our case study. Table 1 shows the requirements for *Booking* and *Payment Release*. These requirements should be fulfilled by the IUT in order to satisfy the service goals. They are added as comments to the model in Figure 5.

Verification. The design models of Holiday Booking (HB) composite REST web service are translated to UPTA with the help of transformation tool. Here, we only show an excerpt of UPTA in Figure 6. The detailed model and the specifications of the partner web services are available in (Rauf et al., 2013).

Table 1: Requirements of Holiday Booking CWS (excerpt).

Req	Sub-Requirements
1- Booking	1.1 - A booking should be paid 1.1.1 - A booking should be paid within 24 hours of the booking. 1.1.2 - If a booking is not paid within 24 hours of the booking, then it is canceled by the system 1.1.3 - A confirmed paid booking, waits for user check in
2- ...	2.1 - ...
4- Payment Release	4.1 - If the user checks in then the payment must be released to the hotel. 4.2 - When the payment is released to the hotel, HB CWS must notify the hotel about release of the payment

The verification properties are specialized for our case study and some of them are mentioned below.

Deadlock Freeness. The HB Service, the hotel service and the payment service models are all deadlock free. This means that the composite service is never reach to a state that cannot preform a transition (i.e., $A[]$ *not deadlock*). Note that the following queries are made for complete model and only some of them can be traced in Figure 6.

Reachability. All the locations in the HB service are reachable. This means that the model receives and sends messages to the partner services smoothly and the model is validated for its basic behavior (i.e., $E\Diamond CompService.r$), where r is the last location in the TA model and indicates that all processes for certain booking is completed.

Safety. Some of the safety properties in our model are: a) Payment should be released iff the user has checked in, i.e., ($E\Box CompService.h2$ imply $CardService.c2$), where $c2$ is the location after check-in and $h2$ is the location after payment release, b) If the payment is released by the HB service then the Hotel service is paid, i.e., ($E\ \Box\ CompService.h2$ imply $HotelService.p$), where p is the location in Hotel service model for hotel payment.

Liveness. Some of the liveness properties in the model are: a) When the payment is not paid within 24 hours, the booking is canceled (i.e., $CompService.c$ and $compService.cl > 24 \rightsquigarrow CompService.b1$), where c indicates waiting for the payment, cl indicates clock of the model and $b1$ indicates the booking request is going to cancel due to the delay, b) If the Hotel Service does not confirm within 3 days then the booking is considered not confirmed (i.e., $CompService.o$ and $CompService.cl > 3 \rightsquigarrow CompService.n$), where o is the location for waiting for the hotel response and n is the location for canceling.

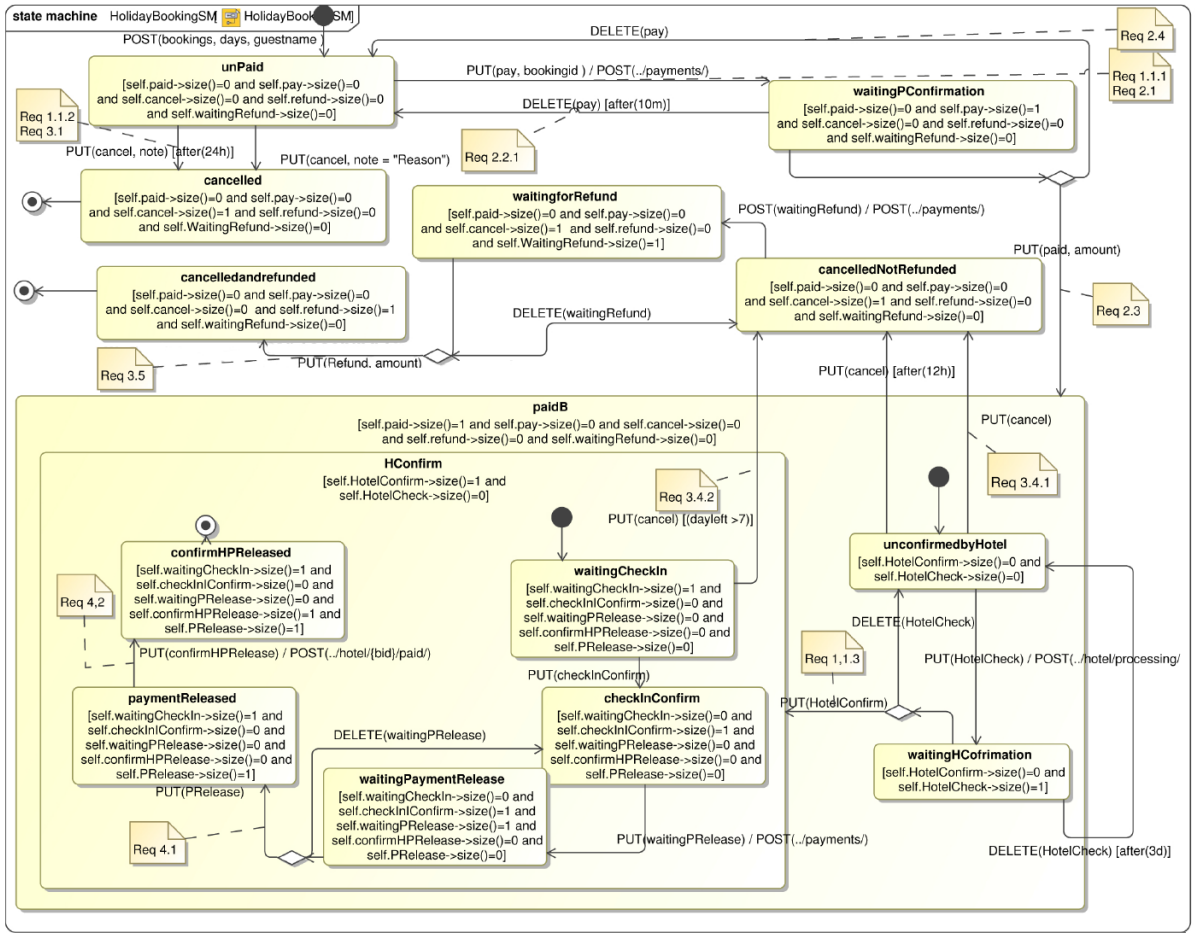


Figure 5: UML State Machine of Holiday Booking Composite REST Web Service.

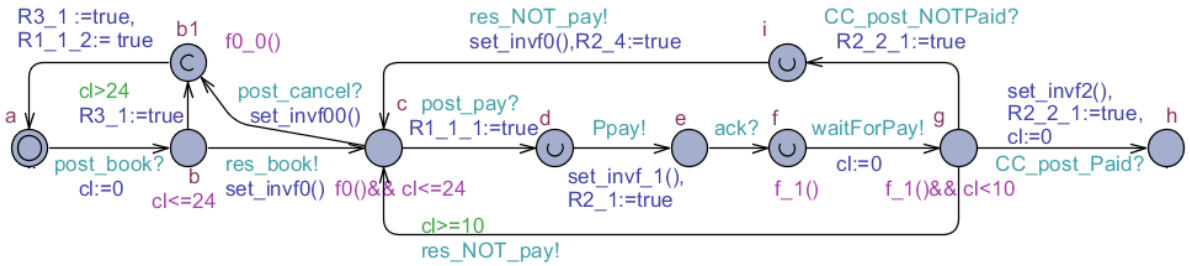


Figure 6: Excerpt of UPTA model of Holiday Booking Composite REST Web Service.

Test Environment. The environment model specifies the user actions, such as booking, canceling a reservation, requesting for the payment, paying, refunding and checking in. These are created from the observable channel synchronizations of the composite web service. The automaton in Figure 7 shows the environment model satisfying edge and requirements coverage. In Figure 7 they are encoded in the guard as a `verdict()` boolean function in the form: $r_1 \&\& \ldots r_n \&\& \&e_1 \ldots \&\& \&e_m$ where r_i and e_j are variables corresponding to requirements and, respectively, to edges of the composite web service in Figure 6. Whenever the verdict function evaluates to TRUE environment model can go to the final location.

Test Setup. Similar to Figure 2, the test setup comprises the TRON engine, the adapter, and the IUT. The IUT is a web service composition of three web services: Holiday Booking, Hotel and Payment Services. The test adapter composed of a set of test cases which satisfy the test requirements that are listed in Table1.

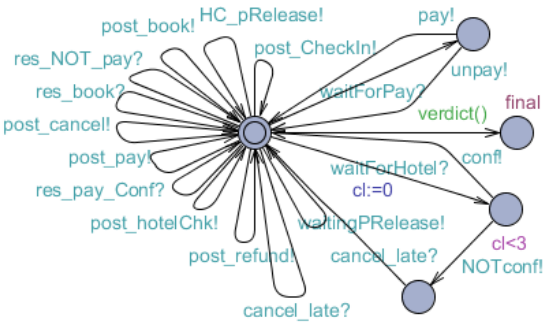


Figure 7: Environment model.

5 EVALUATION

The UML state machines of the HB composite REST web service had 14 states and 25 transitions. These were translated into a UPTA model with 34 locations and 46 edges. Similarly, the state machines of Payment service had 3 states and 4 transitions which transformed in to a UPTA model with 5 locations and 6 edges. The Hotel service had 4 states and 5 transitions that were translated into 7 locations and 9 edges. In addition, the environment model created had 4 locations and 13 edges.

One issue with using formal tools like UPPAAL for verification and test generation, is the scalability of the approach, due to the state space explosion. In contrast to offline test generation, where the entire state space has to be computed, in online test generation only the symbolic states following the current symbolic states have to computed. This reduces drastically the number of symbolic states making the test generation less prone to space explosion and thus more scalable. For instance, the number of explored symbolic states when generating, with the `verifyta` tool, traces satisfying complete edge coverage (i.e., $\&e_1 \ldots \&\&e_m$, where e_j are tracking variables corresponding to all m edges of the HBS models) was 974. In the contrast, the maximum number of symbolic states reported by TRON during a test session achieving complete edge coverage was 12 (see Figure11).

For benchmarking the verification process, we have used the `verifyta` command line utility of UPPAAL for verification of the specified 5 properties. We have used the `memtime` tool to measure the time and memory needed for verification. The result showed in average 0.20 seconds and 54996 KB of memory being used. Although the memory utilization depends heavily on the symbolic state space, it shows that the current size models leave room for scalability of the approach. A known limitation of UPPAAL is that the maximum memory size it can use is close to 4GB due to its 32-bit architecture. Figure 11 plots the evolution of the number of symbolic states for 10000 model time units (20 seconds).

In order to evaluate the efficiency of our approach, we compared the specification coverage with the code coverage yielded by a given test run. Since we had access to the source code of the IUT, we used the *coverage* tool for Python (pyt, 2013) to report the code coverage for each test session. The Table 2 lists results of several measurements:

Table 2: Correspondence between code coverage and edge coverage.

Run	Edge Coverage	Code Coverage
1	64 %	55%
2	80%	67%
3	100%	78%

Although many of the errors were caused by modeling mistakes, testing revealed some errors in the implementation as well. For instance, in the HB service, there was an error in sending *cancel* request and another error found in the POST header in *refund* request. Also in the Hotel service, the confirmation was sent by the wrong method, so it was rejected by Holiday Booking service.

In order to evaluate the fault detection capabilities of our approach, we have manually created 30 mutated versions of the original HB service program code. Each mutation had one fault seeded in the code, for instance replacing POST with DELETE, removing one line of the source code, change of logical conditions, etc. The faults were always seeded in those parts of the code that is covered when achieving 100% edge coverage of the model. We assumed that the original version of the composite web service is the correct one, as we were able to run the 100 test sessions in TRON against it. For each mutated version of the composite web service, we set the TRON to execute 100 test sessions against it. When a fault was discovered, the mutant was considered as *killed*. If the mutated statement has been covered by the test runs but no failure was detected, we mark it as *alive*. Out of the 30 mutated programs, 28 mutants were killed and 2 were alive. This resulted into a mutation score of 93.3%.

6 RELATED WORK

There is already a large body of work on using model checking techniques for validation and verification of web service compositions. Overviews of such works

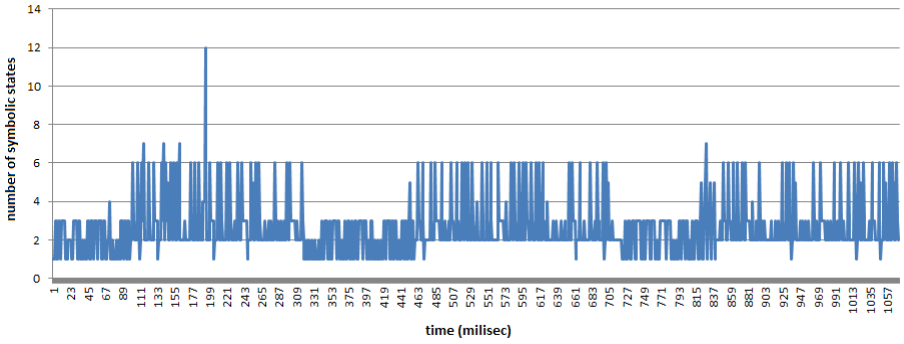


Figure 8: Evolution of symbolic states.

can be found in (Rusli et al., 2011) and (Bozkurt and other, 2010). Mostly authors have used web service specific specification languages as their starting point and converted the specifications to an intermediate language that is accepted by model checking tools. Then, by taking advantage of the model checking tool capabilities they performed simulation, verification or test generation via model-checking. Most of these works use the selected model-checking tool only for simulation and verification; only a handful generate abstract tests from the verification conditions, and in most cases it is not clear how the abstract test cases (i.e., the counterexample traces) are transformed into executable ones and executed. In the following, we will revisit those works which are most similar to ours.

We can distinguish roughly two approach categories: those that target the PROMELA language (Part and Peschke, 2003) which is the input language for the SPIN model-checker (Holzmann, 1997), and those that target the UPPAAL timed automata which is the input language for the UPPAAL model-checker (Behrmann et al.,).

In the first category, the vast majority of approaches have used BPEL or OWL-S(Martin et al., 2004) for the specification of the web service composition. For instance, Garcia (García-Fanjul et al., 2006) generates test cases using test case specifications created from counterexamples that are obtained from model checking. The transition coverage criterion is used to identify transitions in BPEL specification that define the test requirements for producing test cases. These transitions are mapped to the model and expressed in terms of LTL property expressions. Transition coverage is obtained by repeatedly executing the tool with each previously identified transition.

Fu. et al. (Fu et al., 2005) provide framework for analyzing, designing and verifying web service compositions. Their work provides both bottom-up and top-down approach to analyze the interaction between web services. In top-down approach, the desired conversation of a web service is specified as guarded automaton that are converted to PROMELA and used as input to SPIN model-checker. The bottom-up approach translates BPEL to guarded automaton and then used with SPIN model-checker after translating guarded automaton to PROMELA. The web service conversations are analyzed for synchronization in order to verify their compatibility.

One distinct approach is given by Huang et al. (Huang et al., 2005). They automatically translate OWL-S specification of composite web service into a C-like specification language and PDDL through an integrated process. These can be processed with the BLAST model-checker which can generate positive and negative test cases during model checking of a particular formula and test the web service using the test cases.

These works focus on BPEL processes and OWL-S, this makes them dependent on specific execution languages for SOAP based services whereas our work is not dependent on implementation and supports REST architectural style. In addition, their work does not support requirement traceability and is not clear how tests are generated and executed. Furthermore, the works that use the PROMELA language for specification do not address real-time properties, due to the limited support for time in PROMELA.

In the second category, researchers have targeted timed automata specifications. In (Cambronero et al., 2011), Cambronero et al. verify and validate web services choreography by translating a subset of WS-CDL into a network of timed automata and then use UPPAAL tool for validation and verification of the described system. They also capture capture requirements by extending KAOS goal model and implement them. The work is supported by WST tool that provides model transformation of timed composite web services (Cambronero et al., 2012). In (Dıaz et al., 2007), Diaz et. al also provide a translation from WS-BPEL to UPPAAL timed automata. Time properties are specified in WS-BPEL and translated to UPPAAL.

However, requirements are not traced explicitly, while verification and testing are not discussed.

Ibrahim and Al-Ani (Ibrahim and Al Ani, 2013) transform BPEL specification to UPAAAL. The specification includes safety and security non-functional properties which are later formulated into guards in the UPPAAL model which could be similar to our verification of requirements. They do not consider neither real-time properties nor test generation.

In (Guermouche and Godart, 2009), Nawal and Godart deal with checking the compatibility of web service choreography supporting asynchronous timed communications using model checking based approach. They use model-checker UPPAAL and present compatibility checking distinguishing between full and partial compatibility and full incompatibility of web services. Our work is somewhat similar to their work as we support time critical stateful REST webs service compositions using UPPAAL, however, in addition to verification we use UPPAAL with TRON to validate the implementation of the web services.

Zhang (Zhang et al., 2011) suggest the use of the temporal logic XYZ/ADL language (Zhu and Tang, 2003) for specifying web server compositions. They transform the specifications into a timed asynchronous communication model (TACM) which are verified in UPPAAL.

In (Lallali et al., 2008), uses BPEL specifications as a reference specification and transform them to an Intermediate Format (IF) based on timed automata and then propose an algorithm to generate test cases. Similar to our approach, tests are generated via simulation in a custom tool, where the exploration is guided by test purposes. One noticeable difference is that time properties are added manually to the IF specification, while we specify them at UML level.

These works provide approaches to verify and validate the service specifications by checking the properties of interest using UPPAAL tool, however our work, in addition to model checking the properties also performs conformance testing of the service composition via online model-based testing with the TRON tool and provides requirement traceability for non-deterministic systems.

7 CONCLUSION

We have presented an integrated approach to design and validate RESTful composite web services. In our approach, a service can invoke other services and exhibit complex and timed behavior, while still complying with the REST architectural style. We showed how to model the service composition in UML, including time properties. We modeled communicating web services and explicitly define the service invocations and receiving service calls.

We use model checking approach with UPPAAL model-checker to verify and validate our design models. From the verified specification, we generate tests using an online model-based testing tool. The use of online MBT proved beneficial as our system under test exhibits non-deterministic behavior due to concurrency and real-time domain.

With the help of requirements traceability mechanism we traced requirements to UML models and, via the UML→UPTA transformation to timed automata models. Their reachability is verified in UPPAAL and they are used as test goals during test generation. Linking requirements to generated tests allowed us to quickly see which requirements have been validated and which have not. In addition, it allows us to identify from which parts of the specification/implementation the detected error has originated.

We exemplified our approach with a relatively complex case study of a holiday booking web service and we provided preliminary evaluation results.

REFERENCES

(2013). Code coverage measurement for Python – coverage, v. 3.6. Online at https://pypi.python.org/pypi/coverage. retrieved: 20.08.2013.

(2013). Nomagic MagicDraw webpage at http://www.nomagic.com/products/magicdraw/.

Alur, R. et al. (1990). Model-checking for real-time systems. In *Logic in Computer Science, 1990. LICS'90, Proceedings., Fifth Annual IEEE Symposium on e*, pages 414–425. IEEE.

Behrmann, G. et al. Uppaal 4.0. In *QEST '06 Proceedings of the 3rd international conference on the Quantitative Evaluation of Systems*, pages 125 – 126. IEEE Computer Society Washington, DC, USA.

Birgit Demuth, C. W. (2009). Model and Object Verification by Using Dresden OCL. In *Proceedings of the Russian-German Workshop Innovation Information Technologies: Theory and Practice,*, pages 81–89.

Bozkurt, M. and other (2010). Testing web services: A survey. *Department of Computer Science, Kings College London, Tech. Rep. TR-10-01*.

Budgen, D., Burn, A. J., Brereton, O. P., Kitchenham, B. A., and Pretorius, R. (2011). Empirical evidence about the uml: a systematic literature review. *Software: Practice and Experience*, 41(4):363–392.

Cambronero, M. et al. (2012). Wst: a tool supporting timed composite web services model transformation. *Simulation*, 88(3):349–364.

Cambronero, M. E. et al. (2011). Validation and verification of web services choreographies by using timed automata. *Journal of Logic and Algebraic Programming*, 80(1):25–49.

Clarke, E. M., Grumberg, O., and Long, D. E. (1994). Model checking and abstraction. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 16(5):1512–1542.

Dıaz, G. et al. (2007). Model checking techniques applied to the design of web services. *CLEI Electronic Journal*, 10(2).

Fu, X. et al. (2005). Synchronizability of conversations among web services. *Software Engineering, IEEE Transactions on*, 31(12):1042–1055.

García-Fanjul, J. et al. (2006). Generating test cases specifications for BPEL compositions of web services using SPIN. In *International Workshop on Web Services–Modeling and Testing (WS-MaTe 2006)*, page 83.

Guermouche, N. and Godart, C. (2009). Timed model checking based approach for web services analysis. In *Web Services, 2009. ICWS 2009. IEEE International Conference on*, pages 213–221. IEEE.

Hessel, A. et al. (2008). Testing Real-Time systems using UPPAAL. In *Formal Methods and Testing*, pages 77–117. Springer-Verlag.

Holovaty, A. and Kaplan-Moss, J. (2009). *The definitive guide to Django: Web development done right*. Apress.

Holzmann, G. J. (1997). The model checker SPIN. *Software Engineering, IEEE Transactions on*, 23(5):279–295.

Huang, H. et al. (2005). Automated model checking and testing for composite web services. In *Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005. Eighth IEEE International Symposium on*, pages 300–307. IEEE.

Ibrahim, N. and Al Ani, I. (2013). Beyond functional verification of web services compositions. *Journal of Emerging Trends in Computing and Information Sciences*, 4, Special Issue:25–30.

Koskinen, M. et al. (2013). Combining Model-based Testing and Continuous Integration. In *Proceddings of the International Conference on Software Engineering Advances (ICSEA 2013)*. IARIA. TO APPEAR.

Lallali, M. et al. (2008). Automatic timed test case generation for web services composition. In *on Web Services, 2008. ECOWS'08. IEEE Sixth European Conference*, pages 53–62. IEEE.

Larsen, K. G. et al. (1997). UPPAAL in a nutshell. *International Journal on Software Tools for Technology Transfer (STTT)*, 1(1):134–152.

Larsen, K. G. et al. (2009). UPPAAL Tron user manual. *CISS, BRICS, Aalborg University, Aalborg, Denmark*.

Martin, D. et al. (2004). OWL-S: Semantic markup for web services. *W3C member submission*, 22:2007–04.

Nobakht, M. and Truscan, D. (2013). An Approach for Validation, Verification, and Model-Based Testing of UML-Based Real-Time Systems. In Lavazza, L., Oberhauser, R., Martin, A., Hassine, J., Gebhart, M., and Jntti, M., editors, *ICSEA 2013, The Eighth International Conference on Software Engineering Advances*, pages 79–85. IARIA.

Part, I. and Peschke, M. (2003). Design and validation of computer protocols.

Porres, I. and Rauf, I. (2011). Modeling behavioral RESTful web service interfaces in UML. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 1598–1605. ACM.

Rauf, I. et al. (2012). Analyzing Consistency of Behavioral REST Web Service Interfaces. In Silva, J. and Tiezzi, F., editors, *The 8th International Workshop on Automated Specification and Verification of Web Systems*, Electronic Proceedings in Theoretical Computer Science, page 115. EPTCS.

Rauf, I. and Porres, I. (2011). Beyond CRUD. pages 117–135.

Rauf, I., Siavashi, F., Truscan, D., and Porres, I. (2013). An Integrated Approach to Design and Validate REST Web Service Compositions. Technical Report 1097.

Richardson, L. and Ruby, S. (2008). *RESTful web services*. O'Reilly.

Rusli, H. M. et al. (2011). Testing Web services composition: a mapping study. *Communications of the IBIMA*, 2007:34–48.

UML, O. (2009). 2.2 Superstructure Specification. *OMG ed*. http://www.omg.org/spec/UML/2.2/.

Utting, M. and Legeard, B. (2007). *Practical Model-Based Testing: A Tools Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zhang, G. et al. (2011). Model checking for asynchronous web service composition based on xyz/adl. In *Web Information Systems and Mining*, pages 428–435. Springer.

Zhu, X.-Y. and Tang, Z.-S. (2003). A temporal logic-based software architecture description language xyz/adl. *Journal of Software*, 14(4):713–720.

SHORT PAPERS

On Metrics for Measuring Fragmentation of Federation over SPARQL Endpoints

Nur Aini Rakhmawati, Marcel Karnstedt, Michael Hausenblas and Stefan Decker
INSIGHT Centre, National University of Ireland, Galway, Ireland
{*f_author, s_author*}*@deri.org*

Keywords: Linked Data, Data Distribution, Federated SPARQL Query, SPARQL Endpoint.

Abstract: Processing a federated query in Linked Data is challenging because it needs to consider the number of sources, the source locations as well as heterogeneous system such as hardware, software and data structure and distribution. In this work, we investigate the relationship between the data distribution and the communication cost in a federated SPARQL query framework. We introduce the spreading factor as a dataset metric for computing the distribution of classes and properties throughout a set of data sources. To observe the relationship between the spreading factor and the communication cost, we generate 9 datasets by using several data fragmentation and allocation strategies. Our experimental results showed that the spreading factor is correlated with the communication cost between a federated engine and the SPARQL endpoints . In terms of partitioning strategies, partitioning triples based on the properties and classes can minimize the communication cost. However, such partitioning can also reduce the performance of SPARQL endpoint within the federation framework.

1 INTRODUCTION

Processing a federated query in the Linked Data is challenging because it needs to consider the number of the sources, the source locations and heterogeneous system such as the hardware, the software and the data structure and the distribution. A federated SPARQL query can be easily formulated by using the SERVICE keyword. Nevertheless, determining the datasource address that follows SERVICE keywords can be an obstacle in writing a query because prior knowledge data is required. To address this issue, several approaches (Rakhmawati et al., 2013) have been developed with the objective of hiding SERVICE keyword and data sources location from the user. In these approaches, the federated engines receive a query from the user, parse the query into sub queries, decide the location of each sub query and distribute the sub queries to the relevant sources. A sub query can be delivered to more than one data source if the desired answer occurs in the multiple sources. Thus, the distribution of the data can affect the federation performance (Rakhmawati and Hausenblas, 2012). As an example, consider two datasets shown in Figure 1. Each dataset contains a list of personal information using the FOAF(http://xmlns.com/foaf/spec/) vocabulary. If the user asks for the list of all person names, the federated engine must send a query to all data-

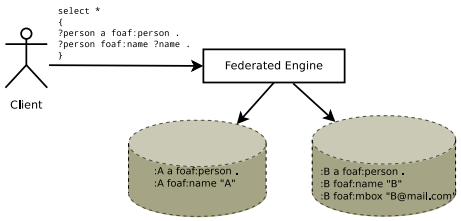


Figure 1: Example of Federated SPARQL Query Involving Many Datasets.

sources. Consequently, the communication cost between the federated engine and data sources would be expensive.

In this study, we investigate the effect of data distribution on the federated engine performance. We propose two composite metrics to calculate the presence of classes and properties across datasets. These metrics can provide insight into the data distribution in the dataset which ultimately, it can determine the communication cost between the federated engine and SPARQL Endpoints. In order to evaluate our metrics, we use several fragmentation and allocation strategies to generate different shapes of data distribution. After that, we run a static query set over those data distributions. Our data distribution strategies could be useful for benchmarking and controlled systems such as organization system, but they can not be address the problem in the federated Linked

Open Data environment because the Linked Data publisher has the power to control the dataset generation. The existing evaluations for assessing the federation over SPARQL endpoints (Montoya et al., 2012; Schwarte et al., 2012) usually run their experiment over different datasets and different query sets. In fact, the performance of the federated engine is influenced by both dataset and query set. As a result, the performance results may vary. For benchmarking, a better comparison of federated engines performance can made with either static query sets over different datasets or static dataset with various query sets.

We only perform our observation on federation over SPARQL endpoints. Query with a SERVICE keyword is also out of the scope of our study because the query only goes to the specified source. In other words, the data distribution does not influence the performance of the federation engine in that query. Our contributions can be stated as follows: 1) We investigate the effects of data fragmentation and allocation on the communication cost of the Federated SPARQL query. 2) We introduce the spreading factor as a metric for calculating the distribution of data across a dataset. In addition, we present the relationship between the spreading factor and the communication cost of federated SPARQL queries. 3) Lastly, we create datasets for evaluating the spreading factor metric drawing from the real datasets. In particular, we provide datasets and a dataset generator that can be useful for benchmarking purpose.

2 RELATED WORKS

Primitive data metrics such as the number of triples, the number of literals are not sufficiently representative to reveal the essential characteristics of the datasets. Thus, Duan (Duan et al., 2011) introduced a structuredness notion. Since this notion is applied to a single RDF repository, it is not suitable for federated SPARQL queries which should consider the data allocation in each repository as well as the number of data sources involved in the dataset.

There are several data partitioning approaches for RDF data clustering repository such as vertical partitioning (Abadi et al., 2007) and Property Table partitioning (Huang et al., 2011). However, the communication in the RDF data clustering is totally different than the communication in the federated SPARQL query. In data clustering, several machines need to communicate with each other in order to execute a query, whereas in the federated SPARQL query, there is no interaction amongst SPARQL endpoints. The mediator has a role to communicate to each SPARQL endpoint during query execution in the federated SPARQL query. Nevertheless, we apply RDF data clustering strategies to generate the datasets for evaluation.

The existing evaluations of the federation frameworks used data partitioning in their experiment by adopting data clustering strategies. Prasser (Prasser et al., 2012) implemented three partitions: naturally-partitioned, horizontally-partitioned and randomly-partitioned. Fedbench(Schmidt et al., 2011) divided the SP2B (Schmidt et al., 2009) dataset into several partitions to run one of their evaluations. Our prior work (Rakhmawati and Hausenblas, 2012) observed the impact of data distribution on federated query execution which particularly focus on the number of sources involved, the number of links and the populated entities in several sources. In this work, we extend our previous evaluation by implementing more data partitioning schemes and we investigate the effect of the distribution of classes and properties throughout the dataset partitions on the performance of federated SPARQL query.

3 SPREADING FACTOR OF DATASET

Federated engines generally use a data catalogue to predict the most relevant sources for a sub query. The data catalogue mostly consists of a list of predicates and classes. Apart from deciding the destination of the sub queries, a data catalogue can help federated engine generate set of query execution plans. Hence, we consider computing the Spreading factor of dataset to analyse the distribution of classes and properties throughout the dataset. We initially define the dataset used in this paper as follows:

Definition 1. *Dataset D is a finite set of data sources d. In the context of federation over SPARQL endpoints, d denotes a set of triple statements t that can be accessed by a SPARQL endpoint. For each SPARQL endpoint, there exists multiple RDF graphs.*

In our work, we ignore the existence of graphs, because we are only interested in the occurrences of properties and classes in the SPARQL endpoint.

Definition 2. *Let U be the set of all URIs, B be the set of all BlankNodes, L be the set of all Literals, then a triple* $t = (s, p, o) \in (U \cup B) \times U \times (U \cup L \cup B)$ *where s is the subject, p is the predicate and o is the object of triple t.*

Later on, we determine the property and the class in the dataset as follows:

Definition 3. *Suppose d is a datasource in the dataset D, then the set $P_d(d,D)$ of properties p in the source d is defined as $P_d(d,D) = \{p|\exists(s,p,o) \in d \wedge d \in D\}$ and the set $P(D)$ of properties p in the dataset D is defined as $P(D) = \{p|p \in P_d(d,D) \wedge d \in D\}$*

Definition 4. *Suppose d is a datasource in the dataset D, then the set $C_d(d,D)$ of classes c in the source d is defined as $C_d(d,D) = \{c|\exists(s,rdftype,c) \in d \wedge d \in D\}$ and the set of classes c in the dataset D is defined as $C(D) = \{c|c \in C_d(d,D) \wedge d \in D\}$*

Given two datasets $D = \{d_1, d_2\}$ as shown in Figure 1. Then $P_d(d_1,D) = \{$rdf:type, foaf:name$\}$, $P_d(d_2,D) = P(D) = \{$rdf:type, foaf:name, foaf:mbox$\}$ and $C_d(d_1,D) = C_d(d_2,D) = C(D) = \{$foaf:person$\}$.

3.1 Spreading Factor of Dataset

With the above definitions of class, property and dataset, now we can describe how we calculate the spreading factor. The spreading factor of the dataset is based on whether or not classes and properties occur. Note that, we do not count the number of times a class and property that are found in the source *d* because the federated engine usually relies on the presence of property in order to predict the data location of a sub query. Given dataset *D* that contains a set of datasets *d*, the normalizing number of occurrences of properties in the Dataset *D* ($OCP(D)$)is calculated as follows: $OCP(D) = \frac{\sum_{\forall d \in D}|P_d(d,D)|}{|P(D)|\times|D|}$ And the normalizing number of occurrences of classes in Dataset *D* ($OCC(D)$) is computed as $OCC(D) = \frac{\sum_{\forall d \in D}|C_d(d,D)|}{|C(D)|\times|D|}$

$OCP(D)$ and $OCC(D)$ have a range value from zero to one. Inspired by the *F-Measure* function, we combine $OCP(D)$ and $OCC(D)$ into a single metric which is called the Spreading Factor $\Gamma(D)$ of the dataset *D*. $\Gamma(D) = \frac{(1+\beta^2)OCP(D)\times OCC(D)}{\beta^2\times OCP(D)+OCC(D)}$ where $\beta = 0.5$

We assign $\beta = 0.5$ in order to put more stress on properties than classes. The intuition is that the highest number of the query pattern delivered to SPARQL endpoint mostly contains constant predicates (Arias et al., 2011). Moreover, the number of distinct properties in the dataset is usually higher than the number of distinct classes in the dataset. The high Γ value indicates that the class and properties are spread out over the dataset.

Look back at our previous example in which we define $P_d(d_1,D)$, $P_d(d_2,D)$, $P(D)$, $C(D)$, $C_d(d_1,D)$,$C_d(d_2,D)$, then we can calculate $OCP(D) = \frac{2+3}{3X2} = 0.833$ and $OCC(D) = \frac{1+1}{1X2} = 1$. Finally, we obtain $\Gamma(D) = 1.172$

3.2 Spreading Factor of Dataset Associated with the Queryset

The spreading factor of a dataset reveals how the whole of classes and properties are distributed over the dataset. However, a query only consists of partial properties and classes in the dataset. Thus, it is necessary to quantify the spreading factor of the dataset with respect to the queryset.

Definition 5. *A query consists of set of triple patterns τ which is formally defined as $\tau(s,p,o) \in (U \cup V) \times (U \cup V) \times (U \cup L \cup V)$ where V is a set of all variables.*

Given a queryset $Q = \{q_1, q_2, \cdots, q_n\}$, the Q-spreading factor γ of dataset *D* associated with queryset *Q* is computed as $\gamma(Q,D) = \sum_{\forall q \in Q} \frac{\sum_{\forall \tau \in q} OC(\tau,D)}{|Q|}$ where the occurrences of class and property for τ is specified as

$$OC(\tau,D) = \begin{cases} \frac{ofD(o_\tau,D)}{|D|} & \text{if } p_\tau \text{ is rdf:type} \\ & \wedge o_\tau \notin V \\ \frac{pfD(p_\tau,D)}{|D|} & \text{if } p_\tau \text{ is not rdf:type} \\ & \wedge p_\tau \notin V \\ \frac{\sum_{\forall d \in D}|P_d(d,D)|}{|D|} & \text{otherwise} \end{cases}$$

$ofD(o,D)$ denotes the occurrences of object *o* in the dataset *D* and $pfD(p,D)$ denotes the occurrences of predicate *p* in the dataset *D* which can be calculated as follows: $ofD(o,D) = \sum_{\forall d \in D} ofd(o,d,D)$ The occurrences of object *o* in the source *d* can be explained as follows:

$$ofd(o,d,D) = \begin{cases} 1 & \text{if } o \in C_d(d,D) \\ 0 & \text{otherwise} \end{cases}$$

$pfD(p,D) = \sum_{\forall d \in D} pfd(p,d,D)$ The occurrence of predicate *p* in the source *d* can be obtained from the following formula:

$$pfd(p,d,D) = \begin{cases} 1 & \text{if } p \in P_d(d,D) \\ 0 & \text{otherwise} \end{cases}$$

Consider an example, given a query and a dataset as shown in Figure 1, then $OC($?person a foaf:person$,D) = 1$ and $OC($?person foaf:name ?name$,D) = 1$ because `foaf:person` and `foaf:name` are located in two data sources. As a result, the q-Spreading factor $\gamma(Q,D)$ is $\frac{1+1}{1} = 2$

4 EVALUATION

We ran our evaluation on an Intel Xeon CPU X5650, 2.67GHz server with Ubuntu Linux 64-bit installed as

Listing 1: Dailymed Sample Triples.

```
dailymeddrug:82 a dailymed:drug
dailymeddrug:82 dailymed:activeingredient dailymeding:
      Phenytoin
dailymeddrug:82 rdfs:label "Dilantin-125 (Suspension)"

dailymeddrug:201 a dailymed:drug
dailymeddrug:201 dailymed:activeingredient dailymeding:
      Ethosuximide
dailymeddrug:201 rdfs:label "Zarontin (Capsule)"

dailymedorg:Parke-Davis a dailymed:organization
dailymedorg:Parke-Davis rdfs:label "Parke-Davis"
dailymedorg:Parke-Davis dailymed:producesDrug
      dailymeddrug:82
dailymedorg:Parke-Davis dailymed:producesDrug
      dailymeddrug:201

dailymeding:Phenytoin a dailymed:ingredients
dailymeding:Phenytoin rdfs:label "Phenytoin"

dailymeding:Ethosuximide a dailymed:ingredients
dailymeding:Ethosuximide rdfs:label "Ethosuximide"
```

the Operating System and Fuseki 1.0 as the SPARQL Endpoint server. For each dataset, we set up Fuseki on different ports. We re-used the query set from our previous work (Rakhmawati and Hausenblas, 2012). We limited the query processing duration to one hour. Each query was executed three times on two federation engines, namely SPLENDID (Görlitz and Staab, 2011) and DARQ (Quilitz and Leser, 2008). These engines were chosen because SPLENDID employs VoID(http://www.w3.org/TR/void/) as data catalogue that contains a list of predicates and entities, while DARQ has a list of predicates which is stored in the Service Description(http://www.w3.org/TR/sparql11-service-description/). Apart from using VoID, SPLENDID also sends a SPARQL ASK query to determine whether or not the source can potentially return the answer. We explain the details of our dataset generation and metrics as follows:

4.1 Data Distribution

To determine the correlation between the communication cost of the federated SPARQL query and the data distribution, we generate 9 datasets by dividing the Dailymed(http://wifo5-03.informatik.uni-mannheim.de/dailymed/) into three partitions based on following strategies:

4.1.1 Graph Partition

Inspired by data clustering for a single RDF storage (Huang et al., 2011), we performed graph partition over our dataset by using METIS (Karypis and Kumar, 1998). The aim of this partition scheme is to reduce the communication needed between machines during the query execution process by storing the connected components of the graph in the same machine. We initially identify the connections of subject and object in different triples. We only consider the URI object which is also a subject in other triples. Intuitively, the reason is that the object which appears as the subject in other triples can create a connection if the triples are located in different dataset partitions. $V(D)$ denotes the set of pairs of subject and object that are connected in the dataset D which can be formally specified as $V(D) = \{(s,o)|\exists s,o,p,p' \in U : (s,p,o) \in D \wedge (o,p',o') \in D'\}$. We assign a numeric identifier for each $s,o \in V(D)$. After that, we create a list of sequential adjacent vertexes for each vertex then uses it as input of METIS API. Run METIS to divide the vertexes and get a list of the partition number of vertexes as output. Finally, we distribute each triple based on the partition number of its subject and object. Consider an example, given Listing 1 as a dataset sample, then

$V(D)$={(dailymeddrug:82, dailymeding:Phenytoin),(dailymeddrug:201, dailymeding:Ethosuximide),(dailymedorg:Parke-Davis, dailymeddrug:82),(dailymedorg:Parke-Davis, dailymeddrug:201)}

Starting an identifier value from one and increment the identifier later, we set the identifier for dailymeddrug:82 = 1, dailymeding:Phenytoin =2, dailymeddrug:201=3, dailymeding:Ethosuximide=4 and dailymedorg:Parke-Davis=5. After that, we can create list of sequential adjacent vertexes $V(D)$ is $\{(2,5),1,(4,5),3,(1,3)\}$. Suppose that we divide the sample of dataset into 2 partitions, then the output of METIS partition is $\{1,1,2,2,1\}$ where each value is the partition number for each vertex. According to the METIS output, we can say that dailymeddrug:82 belongs to partition 1, dailymeding:Phenytoin belongs to partition 1, dailymeddrug:201 belongs to partition 2 and so on. In the end, we have two following partitions:

Partition 1: all triples that contain dailymeddrug:82, dailymeding:Phenytoin and dailymedorg:Parke-Davis

Partition 2: all triples that contain dailymeddrug:201 and dailymeding:Ethosuximide

4.1.2 Entity Partition

The goal of this partition is to distribute the number of entities evenly in each partition. Different classes can be located in a single partition. However, the entities of the same class should be grouped in the same partition until the number of entities reaches the maximum number of entities for each source. We initially create a list of the subjects along with its class ($E(D)$). The set $E(D)$ of pairs of subject and its class in the dataset D is defined as $E(D) = \{(s,o)|\exists(s,rdftype,o) \in D\}$ Then, we sort $E(D)$ by its class o and store each pair

of the subject and object in a partition until the number of pairs of subject and object equals to the total pairs of subject and object divided by the number of partitions. After that, we distribute the remainders of triples in the dataset based on the subject location. Given Listing 1 as a dataset sample, then

$$E(D)=\{(\text{dailymeddrug:82,dailymed:drug}),(\text{dailymeddrug:201},\text{dailymed:drug}),(\text{dailymedorg:Parke-Davis,dailymed:organization}),(\text{dailymeding:Phenytoin,dailymed:ingredients}),(\text{dailymeding:Ethosuximide,dailymed:ingredients})\}$$

Suppose that we split the dataset into two partitions, then the maximum number of entities for each partition is $\frac{|E(D)|}{numberofpartitions} = \frac{5}{2} = 3$ (ceiling 2.5). We place dailymeddrug:82, dailymeddrug:201 and dailymedorg:Parke-Davis in the partition 1 and store the remainders of entities in the partition 2. As the final step, we distribute the related triples based on its subject partition number.

4.1.3 Class Partition

Class Partition divides the dataset based on its classes. The related triples that belong to one entity are placed in the same machine. To begin with, we also create $E(D)$ which was used in Entity partition. Later, we distribute each triple based on the subject class. ike our previous entity partition example, we do the same step to generate $E(D)$. However, in the class partition, we divide the dataset to three partitions since we have three classes (dailymed:drug, dailymed:organization, dailymed:ingredients).

4.1.4 Property Partition

Wilkinson(Wilkinson, 2006) introduced a method for storing RDF data in traditional databases known as Property Table (PT). There are two types of PT partitions: Clustered Property Table and Property-class Table. In our property partition, we do not have a Property class table because we treat all properties in the same manner. We place the triples that have the same property in one data source. Because the number of properties in the dataset is generally high, we allow more than one property to be stored in the same partition as long as we get a balanced number of triples among the partitions. Firstly, we group the triples based on its property. Next, we store each group in a partition until the number of partition triples is less than or equal to the number of dataset triples divided by the number of partitions. For instance, given a dataset as shown in Listing 1, then we have four properties: rdf:type, dailymed:activeingredient, rdf:label and dailymed:producesDrug. Suppose that we want to divide the dataset into 2 partitions, then the maximum number of triples in each partition is $\frac{thenumberoftriples}{thenumberofpartitions} = \frac{14}{2} = 7$. As the following step, we store the triples based on its property as follows: *Partition 1*: five triples with rdf:type property, two triples with dailymed:activeingredient property and *Partition 2*: five triples with rdfs:label property, two triples with dailymed:producesDrug

4.1.5 Triples Partition

The federation framework performance is influenced not only by the federated engine solely, but also depends on the SPARQL Endpoints within the federation framework. In order to keep balanced workload for SPARQL Endpoints, we split up the triples of each source evenly because LUBM (Guo et al., 2005) mentioned that the number of triples can influence the performance of a RDF repository. We created three triple partition datasets (TD, $TD2$, $TD3$). TD is obtained by partitioning the native Dailymed dataset into three parts. $TD2$ and $TD3$ are generated by picking a random starting point within the Dailymed dump file(by picking a random line number).

4.1.6 Hybrid Partition

The Hybrid Partition is a partitioning method that combines two or more previous partition strategies. For instance, if the number of triples in a class is too high, we can distribute the triples to another partition to equalize the number of triples. Since the number of triples in each dataset of the Class Distribution *CD* are not equal, we create *HD* to distribute the triples evenly. However, rdf:type property and rdfs:label property are evenly through all partitions in dataset $HD2$. This distribution is intended for balancing the workload amongst SPARQL Endpoints since those properties are commonly used in our query set.

As shown in those figures, the classes and properties are distributed over most of the partitions in the *GD* dataset. The *PD* has the lowest Spreading Factor among the dataset because each property occurs in exactly one partition and only in one partition has a set of triples that contains rdf:type. The dataset generation code and the generation results can be found at DFedQ github(https://github.com/nurainir/DFedQ).

4.2 Metrics

To calculate the communication cost of the the federated SPARQL query, we compute the data transfer

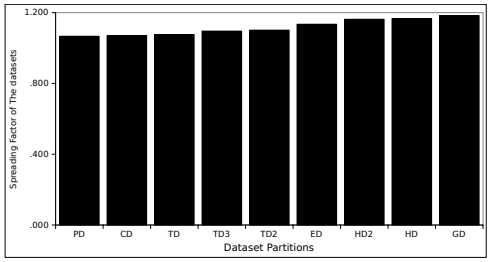


Figure 2: Spreading Factor of Dataset.

volume between the federated engine and SPARQL Endpoints. The data transfer volume includes the amount of data both sent and received by the mediator. Apart from capturing the data transmission, we also measure the requests workload (*RW*) during query execution. *RW* is calculated as $RW = \frac{RQ}{T*SS}$ where *RQ* refers to the number of requests sent by the federated engine to all SPARQL Endpoints, *T* denotes the duration between when a query is received by the federated engine and when its results starts to be dispatched to the client and *SS* is the number of selected sources. Furthermore, we also measure the response time that is required by a federated engine to execute a query.

For the sake of readability, we aggregate each performance metric results into a single value. In order to avoid trade-offs among queries, we assign a weight to each query using the the variable counting strategy from the ARQ Jena (Stocker and Seaborne, 2007). This weight indicate the complexity of the query based on the selectivity of the variable position and the impact of variables on the source selection process. The complexity of query can influence the federation performance. Hence, we normalize each performance metric result by dividing the metric value with the weight of the associated query. In the context of federated SPARQL queries, we set the weight of the predicate variable equals to the weight of the subject variable since most of the federated engines rely on a list of predicates to decide the data location. Note that, a triple pattern can contain more than one variable. The details of the weight of subject variable w_s, predicate variable w_p and object variable w_o for the triple pattern τ can be explained as follows:

$$w_s(\tau)=\begin{cases} 3 & \text{if the subject of triple pattern } \tau \in V \\ 0 & \text{otherwise} \end{cases}$$

$$w_p(\tau)=\begin{cases} 3 & \text{if the predicate of triple pattern } \tau \in V \\ 0 & \text{otherwise} \end{cases}$$

$$w_o(\tau)=\begin{cases} 1 & \text{if the object of triple pattern } \tau \in V \\ 0 & \text{otherwise} \end{cases}$$

Finally, we can compute the weight of query q: $weight(q) = \sum_{\forall \tau \in q} \frac{w_s(\tau)+w_p(\tau)+w_o(\tau)+1}{MAX\_COST}$ where $MAX\_COST = 8$ because if a triple pattern consists of variables that are located in all positions, the weight of the triple pattern is 8(3+3+1+1). By using the weight of a query, we can align the query performance results afterwards. We do not create a composite metric that combines the response time, the request workload and the data transfer, but rather we calculate each performance metric results individually. Given that *Q* is a set of queries *q* in the evaluation and that *m* is a set of performance metric results associated with the queryset *Q*, then the final metric μ for the evaluation is $\mu(Q,m) = \frac{\sum_{\forall q \in Q} \frac{m_q}{weight(q)}}{|Q|}$

For instances, the query in Figure 1 has a weight $= \frac{3+1}{8} + \frac{3+1+1}{8} = 1.125$. Suppose that the volume of data transmission during this query execution is 10 Mb and we only have one query in the queryset, then $\mu(Q,m)$ can be calculated $\frac{\frac{10}{1.125}}{1} = 8.88$Mb.

5 RESULTS AND DISCUSSION

As seen in Figures 3 and 4, the data transmission between DARQ and SPARQL Endpoints is higher than the data transmission between SPLENDID and SPARQL Endpoints. However, Figures 5 and 6 show that the average requests workload in DARQ is less than the average requests workload in SPLENDID.

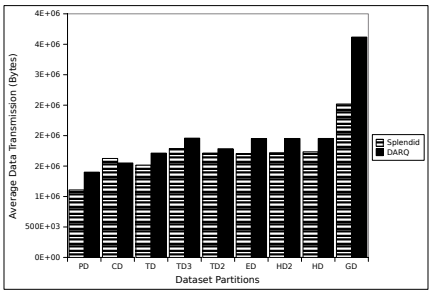


Figure 3: Average Data Transfer Volume Vs the Spreading Factor of Datasets (order by the Spreading Factor value).

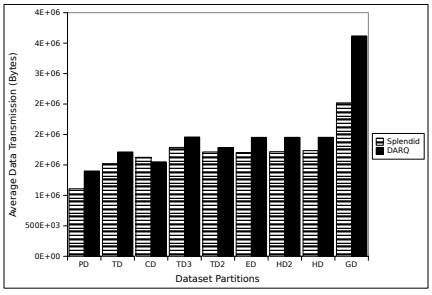


Figure 4: Average Data Transfer Volume Vs the Q-Spreading Factor of Datasets associated with the Queryset(order by the Spreading Factor value).

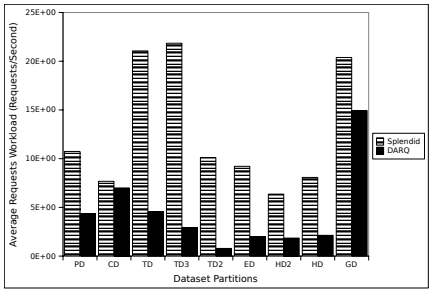


Figure 5: Average Requests Workload Vs the Spreading Factor of Datasets(order by the Spreading Factor value).

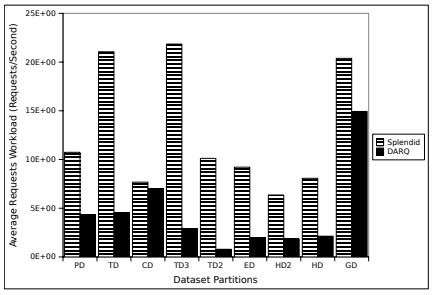


Figure 6: Average Requests Workload Vs the Q-Spreading Factor of Datasets associated with the Queryset(order by the Spreading Factor value).

This is because DARQ never sends SPARQL ASK queries in order to predict the most relevant source for each sub query.

Overall, data transmission increases gradually in line with the Spreading Factor of a dataset. However, the data transmission rises dramatically for *GD* distribution. This indicates that in the context of Federated SPARQL queries, data clustering based on its property and class is better than data clustering based on related entities such as Graph Partition. The reason behind this conclusion is that the source selection in federated query engine depends on classes and properties occurrences. Furthermore, when the federated engines generate query plans, they use optimization techniques based on the statistical predicates and classes.

Although a small Spreading Factor can minimize the communication cost, it can also reduce the SPARQL Endpoint performance. As shown in Figure 5 and 6, a small Spreading Factor can lead to the high number of requests received by SPARQL Endpoint in one second because in the property distribution, the federated engine mostly sends different query patterns to multiple datasource. Moreover, the SPARQL endpoint that stores the popular predicates such as rdf:type and rdfs:label will receive more requests than other SPARQL endpoints. Consequently, this such condition can lead to incomplete results because when overloaded, the SPARQL Endpoint might reject requests (e.g Sindice SPARQL endpoint(http://sindice.com/) only allows one client sending one query per second). Poor performance is also shown at the highest value of Spreading Factor of the dataset (GD) because the entities are spread over the dataset partitions. Hence, with the calculation of the spreading factor of the dataset, the federated engine can create a query optimization which attempts to adapt the dataset characteristic that is shown from the spreading factor value. For instance, if the dataset has too small Spreading Factor, the federated engine should maintain a timer to send several requests to the same SPARQL endpoint in order to keep the sustainability of the SPARQL endpoint as well as avoid the incomplete answer.

6 CONCLUSION

We have implemented various data distribution strategies to partition classes and properties over dataset partitions. We introduced two notions of dataset metrics, namely the Spreading Factor of a dataset and the Spreading Factor of a Dataset associated with the query set. These metrics expose the distribution of classes and properties over the dataset partitions. Our experiment results revealed that the class and property distribution effects on the communication cost between the federated engine and SPARQL endpoints. However, it does not significantly influence the request workload of a SPARQL endpoint. Partitioning triples based on the properties and classes can minimize the communication cost. However, such partitioning can also reduce the performance of SPARQL endpoints within the federation infrastructure. Further, it can also influence the overall performance of federation framework.

In future work, we will apply other dataset partitioning strategies and use more federated query engines which have different characteristics from DARQ and SPLENDID.

ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and Indonesian Directorate General of Higher Education. Thanks to Soheila for a great discussion

REFERENCES

Abadi, D. J., Marcus, A., Madden, S. R., and Hollenbach, K. (2007). Scalable semantic web data management using vertical partitioning. In *Proceedings of the 33rd VLDB*, VLDB '07, pages 411–422. VLDB Endowment.

Arias, M., Fernández, J. D., Martínez-Prieto, M. A., and de la Fuente, P. (2011). An empirical study of real-world sparql queries. *CoRR*, abs/1103.5043.

Duan, S., Kementsietsidis, A., Srinivas, K., and Udrea, O. (2011). Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. In *ACM SIGMOD*.

Görlitz, O. and Staab, S. (2011). SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In *Proceedings of the 2nd International Workshop on COLD*, Bonn, Germany.

Guo, Y., Pan, Z., and Heflin, J. (2005). Lubm: A benchmark for owl knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):158 – 182.

Huang, J., Abadi, D. J., and Ren, K. (2011). Scalable sparql querying of large rdf graphs. *PVLDB*, 4(11):1123–1134.

Karypis, G. and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392.

Montoya, G., Vidal, M.-E., Corcho, Ó., Ruckhaus, E., and Aranda, C. B. (2012). Benchmarking federated sparql query engines: Are existing testbeds enough? In *ISWC(2)*, pages 313–324.

Prasser, F., Kemper, A., and Kuhn, K. A. (2012). Efficient distributed query processing for autonomous rdf databases. EDBT '12, pages 372–383, New York, NY, USA. ACM.

Quilitz, B. and Leser, U. (2008). Querying distributed rdf data sources with sparql. ESWC'08, pages 524–538, Berlin, Heidelberg. Springer-Verlag.

Rakhmawati, N. A. and Hausenblas, M. (2012). On the impact of data distribution in federated sparql queries. In *ICSC 2012*, pages 255 –260.

Rakhmawati, N. A., Umbrich, J., Karnstedt, M., Hasnain, A., and Hausenblas, M. (2013). Querying over federated sparql endpoints - a state of the art survey. *CoRR*, abs/1306.1723.

Schmidt, M., Grlitz, O., Haase, P., Ladwig, G., Schwarte, A., and Tran, T. (2011). Fedbench: A benchmark suite for federated semantic data query processing. In *ISWC*, volume 7031, pages 585–600. Springer.

Schmidt, M., Hornung, T., Lausen, G., and Pinkel, C. (2009). Spˆ 2bench: a sparql performance benchmark. In *ICDE'09.*, pages 222–233. IEEE.

Schwarte, A., Haase, P., Schmidt, M., Hose, K., and Schenkel, R. (2012). An experience report of large scale federations. *CoRR*, abs/1210.5403.

Stocker, M. and Seaborne, A. (2007). Arqo: The architecture for an arq static query optimizer.

Wilkinson, K. (2006). Jena property table implementation. In *In SSWS*.

Delay-based Congestion Control Mechanism for Video Services
Mechanism including Backward Loading and Real-time Modes

Juha Vihervaara and Pekka Loula
Pori Campus, Tampere University of Technology, Pohjoisranta 11, 28100 Pori, Finland
juha.vihervaara@tut.fi, pekka.loula@tut.fi

Keywords: Delay-based Congestion Control, Video-on-demand Service.

Abstract: Currently the congestion control of the Internet is implemented through overprovisioning and TCP protocol. Unfortunately, TCP is not completely suitable for the use of video services. However, recent studies have shown that these video services represent over half of Internet traffic, with a growing trend. There are also arguments against massive overprovisioning. Due to these factors, there is a need to equip video services with proper congestion control. Unfortunately, most of the congestion control mechanisms developed for the use of video services can only offer low priority services or TCP-friendly real-time services. This paper provides a study in which a new delay-based congestion control mechanism is presented. This mechanism can offer congestion control services for both of these service types.

1 INTRODUCTION

Network operators have relied on overprovisioning and TCP congestion control in order to avoid congestion in their networks and to provide an acceptable service level for their customers. Up to the present time this has been reasonable because most of the traffic has been TCP-based. However, video-based services have become a common source of network traffic (Cisco, 2012). As is well known, video-based services have often used the UDP protocol to implement their transport services. Therefore, these video services operate without any kind of congestion control support.

Broadly speaking, based on these changes in Internet traffic types, there are two possibilities for implementing congestion control in the future. First, significant overprovisioning can be used to ensure the proper function of networks. However, nowadays there are some arguments against massive overprovisioning. For example, massive overprovisioning with unnecessary high power consumption is against green Internet ideology (Gupta, 2003). (Huang, 2005) says that the values of the overprovisioning factor of about 5 or even higher are not uncommon in large IP backbones. Self-similar traffic (Wornell, 1996) also sets challenges for overprovisioning. The studies have shown that variable-bit-rate video, which is transmitted over the Internet, is often self-similar (Rikli, 2011).

The second possibility is to use slight overprovisioning and also equip video-based traffic with congestion control. The current solution, where we use congestion control only with TCP traffic, will not be reasonable in the future because TCP traffic will then no longer play the major role (Cisco, 2012). It could be said that TCP and UDP traffic will take on the opposite roles to former years.

There are two different kinds of transfer modes needed among video services. First, there is a need for a backward loading mode where the delay and bandwidth demands are moderate. For example, the whole video can be loaded to the user equipment before it is watched. Also, proxy servers can be used so that videos can reside near customers. In this case, the service provider can load for example films to proxy servers by using the backward loading mode. Secondly, there is also a need for a real-time mode where the delay and bandwidth demands are important. The user may want to watch a live broadcast. The backward loading mode has to work like a low-priority service where the bandwidth is given away to other connections when the load level of the network is high enough. In contrast, the real-time mode always wants its fair share of the bandwidth.

There are many congestion control mechanisms which are suitable for either low priority service or real-time service. But there has been little research work into developing an integrated mechanism

suitable for both cases. The aim of this paper is to study one such kind of integrated congestion control mechanism. In this case, the term integrated means that these two mechanisms are as convergent as possible.

The sending rate can be adjusted based on the packet losses or delays. Both indicators are utilized with the mechanism of this study, but it can be said that the algorithm is more delay-based than loss-based. The reason for emphasizing the more delay-based approach in this study is that it generates a suitable platform for implementing the background loading mode. In this paper, we try to find out if the delay-based approach is suitable for this kind of congestion control mechanism with the help of simulations.

2 RELATED WORK

There are many different kinds of congestion control algorithms which can be used by video services. Some delay-based algorithms and some algorithms for multimedia streaming are presented below. All of them are suitable for a low priority or real-time service. However, none of them are suitable for both service types.

2.1 Delay-based Congestion Control

DUAL (Wang, 1992) is one of the first algorithms in this area. In this algorithm, the optimal queue length is defined and then a corresponding delay in that queue length is used as a congestion signal.

TCP Vegas (Brakmo, 1994) is perhaps the most known delay-based congestion control mechanism. It calculates the amount of data queued on the routing path. This calculation is made with the help of RTT samples and sophisticated calculations. If the queue size is inside the wanted thresholds, the sending rate is not changed. Otherwise, the sending rate is decreased or increased. TCP Vegas does not try to be TCP-friendly because it was designed to be used in place of TCP Reno.

TCP-LP (Kuzmanovic, 2006) offers low priority services as compared to the normal best effort service offered by the Internet. The low priority service is implemented with the help of congestion control. To achieve a low priority service in the presence of TCP traffic, it is necessary to detect an incipient congestion earlier than TCP. TCP-LP measures one-way packet delays and employs a simple delay-threshold based method for the early indication of congestion. TCP-LP decreases its sending rate after the smoothed one-way delay exceeds the threshold.

LEDBAT (Shalunov, 2012) also provides low priority services by using one-way delay measurements to estimate the amount of queued data on the routing path. It increases or decreases its sending rate based on the predetermined target. Sending rates are increased and decreased more aggressively if the queuing delays are far from the target.

CDG (Hayes, 2011) is a delay-gradient based congestion control algorithm which no longer requires knowledge of the path-specific minimum delay thresholds for the congestion notification. In fact, CDG does not use delay gradients as a main source of congestion discovery. Instead, delay gradients are used to assist the normal TCP Reno-like congestion control to make the right decisions. Gradients are used to differentiate between congestion and non-congestion losses.

2.2 Congestion Control for Multimedia Streaming

Research work has been very active in this area during the last few decades. There are many proposals for multimedia streaming. One approach is to think that without changes TCP is not suitable for streaming media because it uses retransmissions for reliable communication with head of the line blocking. Therefore, the easiest way to implement congestion control oriented TCP-friendly transport layer protocol for multimedia streaming is to emulate the behaviour of TCP but without TCP's reliability. RAP (Rejaie, 1998) and TCP-RMP (Liang, 2002) can be classified as members of this group. TCP-RMP also does smoothing operations for its sending rate because big rate fluctuations with multimedia streaming are undesirable.

The most known proposal for multimedia streaming is DCCP (Kohler, 2006) and its TCP Friendly Rate Control version, abbreviated as TFRC (Floyd, 2008). The main point with DCCP's congestion control is that congestion control is not a part of DCCP itself. Instead, DCCP allows applications to choose from a set of congestion control mechanisms. At this moment, the main mechanism for multimedia streaming is TFRC. TFRC is designed for applications that need a smooth rate, and many real-time multimedia applications are this type. TFRC is an equation-based congestion control mechanism, which uses TCP's throughput equation (Padhye, 1998) to calculate the allowed sending rate.

Google Congestion Control for Real-Time Communication on the World Wide Web (Lundin, 2012) is a new proposal for this area. It uses delay gradients in a sophisticated way to detect congestion so that different frame sizes are taken into account.

3 CONGESTION CONTROL MECHANISM FOR VIDEO SERVICES

In this section, a new congestion control mechanism is presented. This new congestion control mechanism is called Congestion control for VIdeo to Home Internet Service, or CVIHIS.

3.1 Basic Functions

There are some basic choices for CVIHIS. Typically, this kind of service is implemented in a client-server manner because one video server is serving a great number of clients. Therefore, the natural choice is that CVIHIS is a receiver-based mechanism so that all possible calculations are placed on the receiver side.

The real-time mode wants to get its fair share from the network bandwidth. This means that this mode should be TCP-friendly. However, as usual with this kind of video application, the bandwidth should be used in a much smoother way than TCP's congestion control does. Of course, this new mechanism should be stable and scalable which are typical requirements for all congestion control mechanisms. CVIHIS is an end-to-end mechanism because the key element for scalability is the end-to-end argument described in (Saltzer, 1984). The idea behind this argument is that if you want the mechanism to have scalability, the complex issues should be placed at the network endpoints.

The sending rate can be controlled in a rate-based or window-based manner. A natural choice for CVIHIS, and traditionally a more commonly used approach with multimedia streaming, is the rate-based approach. (Akan, 2004) says that the window-based approach is not suitable for continuous multimedia streaming since its ACK-controlled packet injection method does not maintain smooth rate variation.

If the network does not support explicit congestion feedbacks, the sending rate can only be adjusted based on the packet losses or delays. Both indicators are utilized by CVIHIS, but it can be said that the algorithm is more delay-based than loss-based. The reason to emphasize a more delay-based approach is that it generates a suitable platform for implementing the background loading mode. For example, LEBDAT and TCP-LP offer mechanisms so that low–priority applications can withdraw from using bandwidth after the queuing delay exceeds the predefined target.

CVIHIS uses one-way delays for delay measurements instead of round-trip times. The main motivation behind the choice of one-way delay is that this means the necessary calculations can be made on the receiver side. However, there is a drawback with this one-way delay based mode. The clocks of the end-points have to be synchronized with some level of accuracy because we are measuring one-way delays.

3.2 Backward Loading Mode

The rate adaptation schema of CVIHIS is presented in Figure 1. Based on delay measurements, the minimum and maximum delays are picked up. The minDelay value presents a situation in which the queues of the connection path are empty. The minDelay value only includes propagation delay components. It could be said that the minDelay value is the shortest one-way delay value experienced during the lifetime of the connection.

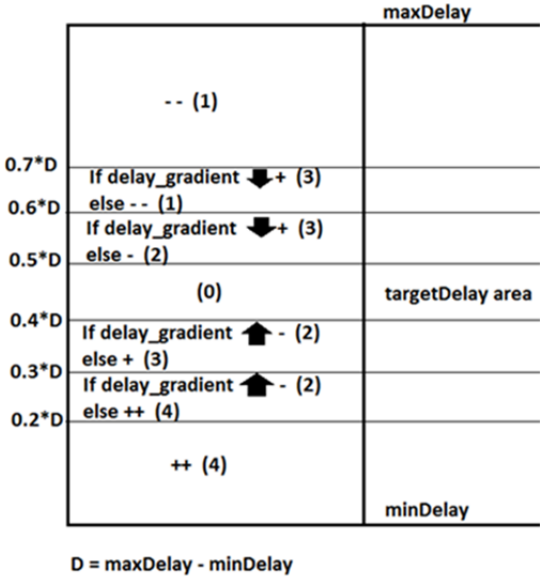


Figure 1: Rate adaptation schema of CVIHIS.

The maxDelay point is found when a packet drop occurs. The delay value of the last received packet before the dropped packet is used for the maxDelay value. In this study, the maxDelay value is totally

updated after each packet drop. However, there can be different kinds of congestion situations in real networks, and these situations generate different kinds of delay values for packet drop situations. Thus, for example, the Exponentially Weighted Moving Average (EWMA) could be used in the calculation of the maximum delay value. The EWMA process would average different kinds of congestion situations.

With the help of the minimum and maximum delay values, the delay space is divided into seven rate adaptation areas. We could also say that the queue of the router is divided into several parts. The centremost delay area is called the targetDelay area. The basic idea is that CVIHIS tries to keep the queue level of the router at the level of this target. Note that this target delay may be over the delay demand of the application. If applications have delay and jitter demands, these demands have to be satisfied by Quality-of-Service mechanisms. CVIHIS is a pure congestion control mechanism, not a Quality-of- Service mechanism.

As shown in Figure 1, the target delay area is not put in the middle of the delay space. Instead, this area is shifted downwards a little so that queuing delays can be kept short. The shiftng used in this case is perhaps too modest for the normal situation. But if the routers are equipped with active queue management mechanisms, this shiftng may be appropriate. The positioning factors for each delay area are presented on the right side of this figure.

There are black arrows inside some delay areas. These black arrows represent the direction of the delay. Thus, delay gradients are also used by CVIHIS for rate adaptation. With the help of delay gradients the target delay area can be achieved without significant oscillation. Using only delay gradients for rate adaptation is not recommended because delay gradients are often too sensitive to second-order delay fluctuations (Wang, 1992).

These delay gradients are obtained by comparing the delay values of two consecutive packets. If the arrow is upwards, the delays are increasing and the delay gradient is positive. This also means that the queue is filling up. If the arrow is downwards, the delays are decreasing and the delay gradient is negative. Therefore, the queue is empting.

As can be seen, we have four delay areas with a black delay gradient arrow. Inside these four areas, the rate adaptation command is based on the actual delay value and the value of the delay gradient. The delay value assigns a certain delay area. Inside this delay area, the delay gradient defines the actual rate adaptation feedback. The two extreme delay areas do not use delay gradients for rate adaptation decisions.

This rate adaptation schema tries to achieve two target values. It tries to drive the queue level to the level of the target delay area. This is done with the help of the actual delay values. The adaptation schema also tries to adapt the sending rate to the level of the bottleneck capacity. This is done with the help of the delay gradients. Based on these two targets, one extra rule can be defined. If there is a conflict of interests between these two targets, the sending rate target is favoured.

CVIHIS can adjust its sending rate through five adjustment steps. Four of these steps are presented in Figure 1 by ++, +, -, --. There are two different level steps for both increasing and decreasing the sending rate. These four adjustment steps are quite moderate so that some kind of additive increase and additive decrease behaviour is achieved. The fifth adjustment step is a multiplicative decrease step which is used after a packet drop. During one round-trip time only one multiplicative decrease step is taken. To simplify matters, we can say that CVIHIS imitates AIMD behaviour (Chiu, 1989).

As stated above, the aim is for CVIHIS to be a receiver-based congestion control mechanism. Therefore all conclusion procedures presented in Figure 1 are implemented on the receiver side. Only the rate adaptation command, which is an integer number, is transmitted to the sender. The values of the commands are presented inside parentheses in Figure 1. The sender increases or decreases its sending rate based on these commands if the measured delay is outside the target delay area.

3.3 Real-time Mode

The backward loading mode backs off when it competes with TCP. This means that we must modify the backward loading code so that it will behave in a more aggressive way to be suitable for the real-time mode. We will take an approach in which we push the minimum delay value upwards in a continuous manner. This approach means that the delay areas, which are presented in Figure 1, are also pushed upwards and therefore will CVIHIS behave more aggressively.

As a matter of fact, CVIHIS competes harder and harder during the pushing phase because the delay areas move higher and higher all the time. This upwards pushing is only done when the competing behaviour is really needed. If the last measured delay value is smaller than the pushed minimum delay value, the minimum delay value is set to the

last measured delay value. With this choice, there is no need to change the code of the sender side. In addition, only two code lines have to be added to the code of the receiver side.

This pushing function makes it possible for CVIHIS to compete against the TCP protocol. But this pushing function also gives other benefits. It alleviates the rerouting problem which is a typical problem for delay-based congestion control mechanisms. It also helps the clock synchronization problem.

In fact, the clock offset is not a problem for CVIHIS. CVIHIS divides the area between the values of the minDelay and maxDelay into several delay areas. CVIHIS can do this correctly if the maxDelay is greater than the minDelay. This means that these delay values can also be negative due to the clock offset. Instead of, clock drift can cause problems. If the delay trend is growing, the minDelay will become outdated. So, it is desirable to update the minDelay value from time to time. In this way, the accumulation of clock drift can be removed. Due to this clock drift problem, it is also perhaps clever to equip the backward loading mode with this pushing function in the future. With the backward loading mode, a small pushing factor could be used. The simulation in this study of the backward loading mode did not include this pushing.

The minimum delay value is pushed upwards by the pushing factor so that the current value is multiplied by this factor. Through simulation it was found that a pushing factor of 1.10 is the value of the suitable parameter group. This 1.10 is quite a high value. This high value indicates that CVIHIS must behave in quite an aggressive manner to be TCP-friendly. In fact, (Yang, 2001) shows that TCP is a quick protocol to utilize extra bandwidth.

4 SIMULATION RESULTS

The simulation program used in this work is NS-2 (Fall, 2013). The congestion control mechanism developed in this work is implemented as a new transport layer protocol of NS-2. The new network component objects have been written using C++-language. These new objects have been added to the NS-2 environment.

4.1 Test Network

The structure of test network is presented in Figure 2. There are six end nodes and two core nodes. Nodes 2 and 3 are core nodes. Because we are dealing with the TCP/IP protocol suite, these core nodes present IP routers. In node 2, the queue size of the bottleneck link is 20 packets. Other nodes are end nodes. For example, CVIHIS's sender resides on node 0 and CVIHIS's receiver on node 4. Some end nodes are represented by broken lines because these nodes are present only if three user connections are needed. In the simulation results, CVIHIS will be depicted in red. TCP connections will be depicted in green and blue.

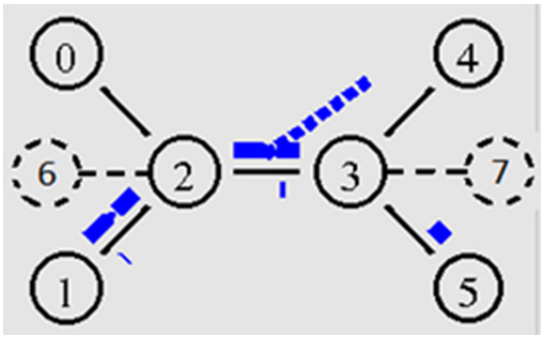


Figure 2: Structure of the test network.

4.2 Backward Loading Mode

Figure 3 presents the results of the simulation in which the back off behaviour of CVIHIS is tested.

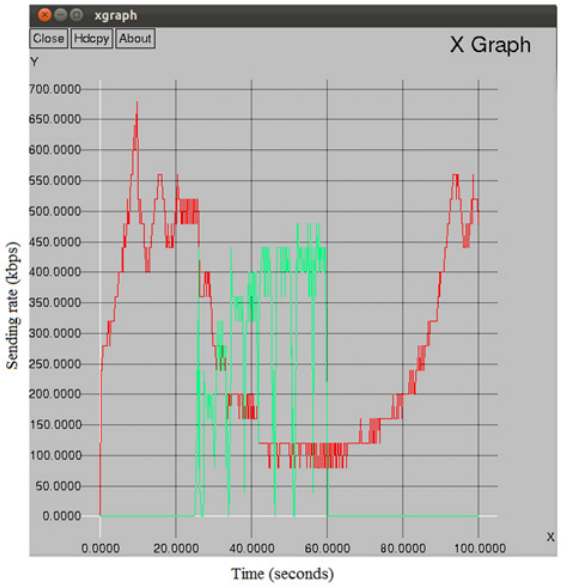


Figure 3: Simulation results of the backward loading.

The CVIHIS connection is started at the beginning of the simulation. The sending rate of CVIHIS is expected to settle to the capacity of the bottleneck link. This initial phase takes about 20 seconds

because no special start phase algorithm is used. After this settling phase, the TCP connection is started. The TCP connection is stopped after the CVIHIS back off action. After that it is checked that CVIHIS increases its sending rate because there is now free capacity in the network.

As can be seen, there is some oscillation behaviour at the beginning of the connection. The first high rate peak is necessary because the CVIHIS queue level probing does not switch on until after the first packet drop. The maxDelay value is not defined before this first drop. The stabilization happens after 20 seconds. As can be seen, the rate also varies after this point. However, this rate variation is generated only by the phase effect of the analysis tool. In order to calculate the sending rates, the packets are grouped. This grouping interval is 0.2 seconds, and it is often not compatible with the packet interval times. As a result, there may be one packet differences between the groups.

This simulation shows that the backward loading mode of CVIHIS works as expected and desired.

4.3 Real-time Mode

The TCP-friendliness of CVIHIS is tested in this section. Tests are made against the TCP Reno version. In fact, there is more behind this phase because the correct rate adjusting parameters for CVIHIS are searched for at the same time. These parameters can be used for tuning CVIHIS to be TCP-friendly. This suitable parameter group has also been used in the simulation case of the backward loading mode.

This parameter group is presented in Table 1. In fact, we are adjusting the sending gaps of the packets. So if we want to decrease the sending rate, the value of the parameter must be over 1. The four leftmost columns present values for the delay areas presented in Figure 1. MD is the multiplicative decrease factor which is used after packet drops. The value of the pushing factor used by the real-time mode is 1.10. Every data packet generates a rate adjustment command. However, after the multiplicative decrease step it takes one round-trip time before the rate adjustment continues.

Table 1: Adjustment parameters of CVIHIS.

++	+	--	-	MD
0.9976	0.999	1.002	1.001	1.23

In the first simulation case, CVIHIS and TCP start with the initial rate of 300 kbps. The capacity of the bottleneck link is 600 kbps. Both connections start at the same time. The simulation result for this case is depicted in Figure 4. In this simulation case, CVIHIS manages very well. It takes almost exactly its fair share from the capacity of the bottleneck. We have also made other simulations in which the initial rates of CVIHIS and TCP have not been equal. In those cases, CVIHIS managed at an acceptable level but not as perfectly as in this first case.

CVIHIS has also been tested against two TCP connections. The CVIHIS connection and the other TCP connection are sending all the time. Another TCP connection is started and stopped during the simulation. The capacity of the bottleneck link is still 600 kbps.

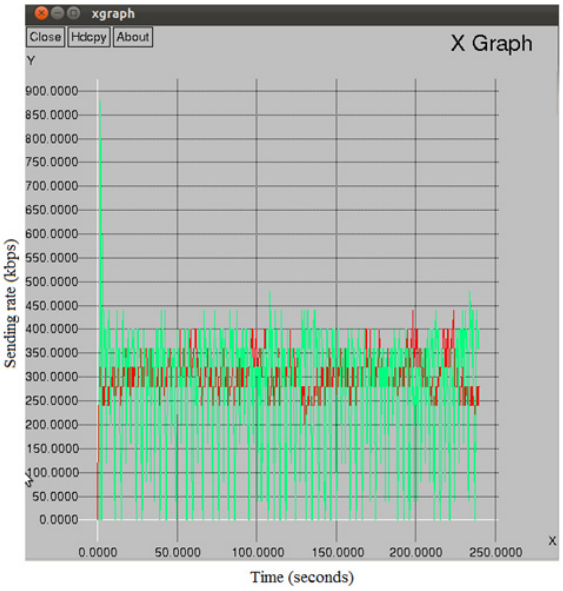


Figure 4: Test against one TCP connection.

In this simulation, it is first investigated if CVIHIS reduces its sending rate after the other TCP connection is started. It is also examined whether CVISIS would increase its sending rate after the other TCP connections is stopped. The simulation result is presented in Figure 5. In this case, CVIHIS manages moderately. It can decrease and increase its rate as required. However, when there are two active TCP connections, CVIHIS exhibits back-off behaviour. The sending rate of 100 kbps is CVIHIS's minimum rate. This moderate minimum rate is used because real-time video services need it.

4.4 Rerouting Problem

One simulation has been done to test the rerouting properties of CVIHIS. The rerouting discovery ability of CVIHIS is based on the pushing of the

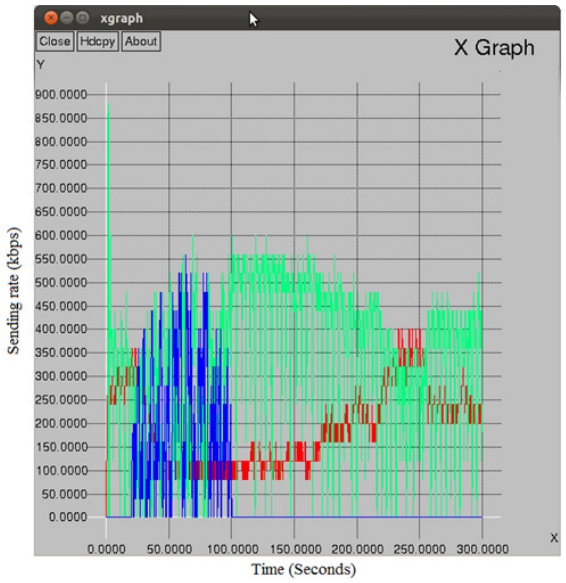


Figure 5: Test against two TCP connections.

minimum delay value and updating of the maximum delay value.

The network layout has been changed for this simulation. There are two possible routes between nodes 2 and 3. There is a direct default route and a backup route. The default route is switched off twiceo times during the simulation. The capacity of the default route is 700 kbps and the capacity of the backup route is 400 kbps. The simulation result is presented in Figure 6. As can be seen, CVIHIS can observe route changes and it can accommodate its sending rate according to the new route.

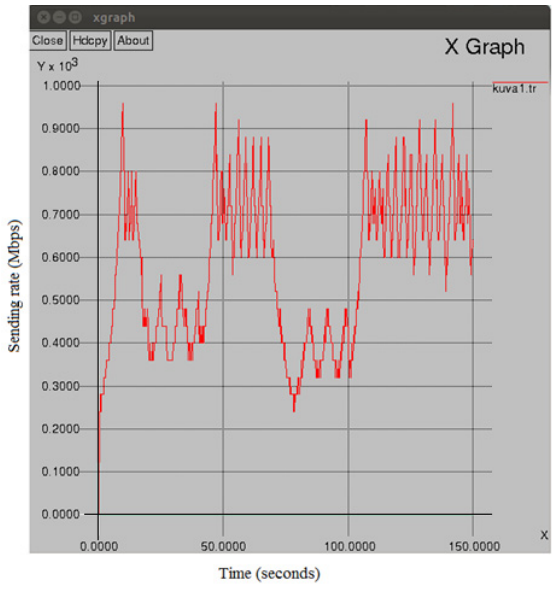


Figure 6: Testing rerouting properties of CVIHIS.

5 CONCLUSIONS

With the help of simulation it was found that this kind of delay-based congestion control approach could be suitable for video services. CVIHIS managed in an acceptable way.

However, it was found that CVIHIS behaves in quite a parameter sensitive way in terms of TCP-friendliness. Small changes to the values of the adjusting parameters caused changes to the TCP-friendliness. This kind of parameter sensitiveness can be problematic. It can also mean that minor changes in the network environment generate problems for TCP-friendliness. Therefore, it is also a good idea to simulate CVIHIS against other TCP versions, like TCP New Reno.

On the other hand, these kinds of findings come as no big surprise. It is a delicate task to make different kinds of mechanisms work together in a harmonious way. TCP and CVIHIS are very different kinds of congestion control mechanisms. TCP is a loss-based mechanism and CVIHIS is merely a delay-based mechanism. TCP is a window-based mechanism and CVIHIS is a rate-based mechanism. Widmer's paper (Widmer, 2001) analyses different kinds of congestion control mechanisms. This paper shows that rate-based congestion control mechanisms have some trouble with TCP-friendliness because TCP uses window-based rate control. Consequently, extra simulations are needed to prove the usefulness of this CVIHIS approach on a large scale.

Some elaboration work is also needed. For example, due to the pushing operation of the minimum delay value, the sending rate of the real-time mode fluctuates. This can be seen in Figure 6. Therefore, some smoothing function could be added in future work. For example, some filtering procedure could be put on the top of the current algorithm to make a low pass filtering procedure.

Instead of specifying a complete protocol, the aim of this study was to specify only a congestion control mechanism that could be used in different parts of the protocol stack. Only a basic mechanism was defined and all kinds of special cases were left outside this presentation. Therefore, all kinds of special cases, such as the start phase of the connection, need extra study in the future.

One possible point for future research could be to test the real-time mode of CVIHIS by using another kind of congestion control approach. This other approach could be a loss-based approach so that the behaviour of TCP congestion control is emulated on the receiver side.

REFERENCES

Akan Ö. 2004. *On the throughput analysis of rate-based and window-based congestion control schemes*. Computer Networks 44 (2004). Pages 701-711.

Brakmo L, O'Malley S, Peterson L. 1994. *TCP Vegas: New Techniques for Congestion Detection and Avoidance*. ACM SIGCOMM London, UK. Pages 24-35.

Cisco. 2012. *Cisco Visual Networking Index: Forecast and Methodology*, 2011-2016. Retrieved 2.2.2013 from http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.

Chiu D, Jain R. 1989. *Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks*. Computer Networks and ISDN Systems 17/1989 Pages 1-14.

Gupta M, Singh S. 2003 *Greening of the Internet*. ACM SIGCOMM, Karlsruhe, Germany.

Hayes D, Armitage G. 2011 *Revisiting TCP Congestion Control using Delay Gradient*. The proceedings of NETWORKING 2011 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011.

Huang Y, Guerin R. 2005. *Does Over-Provisioning Become More or Less Efficient as Networks Grow Larger?*. Proceeding of the 13th IEEE International Conference on Network Protocols 2005.

Fall K. 2013. *A collaboratoion between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC. NS-2 Manual.* www.isi.edu/nsnam/ns/doc/index.html.

Floyd S, Handley M, Padhye J, Widmer J. 2008. *TCP Friendly Rate Control (TFRC): Protocol Specification*. RFC5348.

Kohler E., Handley M., Floyd S., 2006. *Datagram Congestion Control Protocol (DCCP)*. RFC4340.

Kuzmanovic A, Knightly E. 2006. *TCP-LP: low-priority service via end-point congestion control*. IEEE/ACM Trans. Netw. 14(4) August 2006. Pages 739–752.

Liang S, Cheriton D. 2002. *TCP-RTM: Using TCP for Real Time Applications.* IEEE International Conference of Network Protocols '02.

Lundin H, Holmer S,. Alvestrand H. 2012. *"A Google Congestion Control Algorithm for Real-Time Communication on the World Wide Web"*. IETF Informational Draft, Version 3.

Padhye J, Firoiu V, Towsley D, Kurose J, 1998. *"Modeling TCP Throughput: A Simple Model and its Empirical Validation",* Proc ACM SIGCOMM 1998.

Rejaie R, Handley M, Estrin D. 1998. *Architectural considerations for playback of quality adaptive video over the internet.* Technical report 98-686, CS-USC, November 1998.

Rikli N. 2011. *Self-similarity and stationarity of increments in VBR video.* King Saud University Journal of King Saud University – Computer and Information Sciences 4/2012 Pages 7-16, Retrieved 1.2.2013 from www.ksu.ede.sa.

Saltzer J, Reed D, Clark D. 1984. *End- to- end arguments in system design*. ACM Transactions on Computer Systems 2(4). Pages 277-288.

Shalunov S., Hazel G, Iyengar J, Kuehlewind M. 2012.. *Low Extra Delay Background Transport (LEDBAT)*. RFC6817.

Wang, Z., Crowcroft, J. 1992. *Eliminating periodic packet losses in the 4.3-Tahoe BSD TCP congestion control algorithm.* SIGCOMM Computer Communication. Rev. 22(2). Pages 9–16.

Widmer H, Denda R, Mauv M. 2001. *A Survey on TCP-Friendly Congestion Control.* IEEE Network May 2001, Volume:15 Issue 3,. Pages 28-37.

Wornell G. 1996. *Signal processing with fractals: A Wavelet Based Approach*. Prentice Hall. 176 Pages.

Yang Y, Kim M, Lam S. 2001. *Transient Behaviors of TCP-friendly Congestion Control Protocols.* INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Pages 1716 - 1725, vol.3.

An Approach to Transform Public Administration into SOA-based Organizations

J. Sedeño[1,2], C. J. Torrecilla-Salinas[1], M. J. Escalona[1] and M. Mejías[1]
[1]*IWT2 Group. Escuela Técnica Superior de Ingeniería Informática, University of Seville, Seville, Spain*
[2]*Agencia Andaluza de Instituciones Culturales, Junta de Andalucía, Seville, Spain*
{jorge.sedeno, carlos.torrecilla}@iwt2.org, {risoto, mjescalona}@us.es

Keywords: Governance, SOA, Public Administration, e-Government.

Abstract: Nowadays, Service-Oriented Architectures (SOA) is widely spread in private organizations. However, when transferring this knowledge to Public Administration, it is realized that it has not been transformed in terms of its legal nature into organizations capable to operate under the SOA paradigm. This fact prevents public administration bodies from offering the efficient services they have been provided by different boards of governments. A high-level framework to perform this transformation is proposed. Taking it as starting point, an instance of a SOA Target Meta-Model can be obtained by means of an iterative and incremental process based on the analysis of imperatives and focused on the particular business context of each local public administration. This paper briefly presents a practical experience consisting in applying this process to a Spanish regional public administration.

1 INTRODUCTION

Nowadays, SOA is widely spread in private organizations environments both in practice and theory. However, when we want to transfer this body of knowledge to Public Administration, implementations of the SOA paradigm are commonly reduced to implementations of Web Services, orchestration of processes or governance only based on rules of technical character that prevent public organizations from supplying the different public services efficiently and effectively.

Regarding the experience of implementing a SOA architecture and government in the Ministry of Education, Culture and Sport of the Andalusia Regional Government, which is a leading project which applies the usual methodologies of SOA modelling (Arsanjani, 2004) (Marks, 2008) to Public Administration (Sedeño, 2013), we have observed that an alignment with the Public Administration business is required to optimize the aforementioned administration's medium and long-term profits.

We propose a high-level framework to drive this transformation throughout the organization, covering all aspects related to the Public Administration business context. These operations sequences are initiated when certain events take place within this context.

A key point in this process are imperatives-understood either as duty or inexcusable requirements-, than can be defined for Public Administration as those that being overcome, may allow reaching excellence in providing citizens with services that inherently correspond to them, as well as in managing internal processes supporting them.

2 SOA PUBLIC BACKGROUND

The implementation of SOA is a complex technological, organizational and business undertaking. It requires being aware of the process theory and knowing it, as well as a deep understanding of the processes within the organization.

However, implementations in Public Administration demand cognitive and practical studies. There are major differences between private organizations and Public Administration. Therefore, solutions successfully used in private environments, could not apply to Public Administration**.**

This paper deals with providing an approach to transform Public Administration into SOA-based organizations. Thus, it is necessary to understand the business context differences between a public and a private organization.

Unlike private organizations, focused on their profitability and stakeholders' value, public administration bodies have the public interest at heart. Their success goes along with their reaching of social goals. Simultaneously, these organizations have to cope with social and political demands, which are not the regard of the private sector. To compound the problem, public administration bodies create policies or policy advises which are more difficult to identify than private sector organizations' physical output or financial results (Tregear and Jenkins, 2007).

Processes in Public Administration are more complex than in private organizations (Repa, 2006). This is due to the fact that Public Administration is a closely set organizational structure, where processes are formalized and decision-making is slow. This is compounded by the decision-making independence of individual departments, which often pursue their own goals regardless of the whole board of government. The process flow across departments is much more complex and depends on the task scope of individual employees. In addition, all government actions are hedged around with a great number of legal regulations. These make the modification or improvement of processes a hard task to perform. They often require the introduction of changes in the law, which is a lengthy process.

Hence, creating an efficient Administration is impeded by factors (Hall, 2007) such as:

- Increasing bureaucracy (larger number of documents, statements and reports. Spain is organized into four administration levels).
- Complex and frequently changing administrative procedures (Each Administration in our country rules its own laws by adapting official central laws).
- Law inflation (establishment of new laws and regulations by the legislators. In Spain several Administrations duplicate their competencies, either at local or regional level).
- Competition for financial resources (budget item justification, application for EU funds).

Public Administration uses SOA solutions based on web services implantations (Goudos el al, 2007), policies orchestration (Popescu et al, 2012), or business rule compliance (Castellano et al, 2005) without the aim of becoming a reference model to them (Galster et al, 2013), so that it can overcome these barriers and become more competitive.

In contrast, there are methods based on understanding the organizational elements and their relationships and, also, the drive goals analysis (Delgado et al, 2013). In the same way there are several approaches about IT government such as ITIL, COBIT or TOGAF interrelated to the delivery of service. However, these methods do not specify how to transform Public Administrations' nature that is the aim of our paper -the transformation of the Public Administrations in terms of its legal nature into organizations capable to operate under the SOA paradigm-.

Nevertheless, the own administration must be transformed to operate in SOA environments incorporating its own legal and human barriers as part of the methodology. Many Enterprise Architecture (EA) Frameworks for private organizations about reference models based on SOA (Ruas de Oliveira et al, 2010) (Alwadain et al, 2013) are gathered after a literature review, whereas little information about critical success factors (CFS) in Public Administrations is found (Abdul-Manan & Hyland, 2013) (Koumaditis et al, 2013). EA implementation involves products and services that are both in business and industry domains, the technical infrastructure based on open system building blocks, the definition of EA that includes business process architecture, the application architecture and data and technology architectures (McSweeney, 2000). In fact, we have not achieved any generic framework designed specifically to support any kind of public environment.

All concepts exposed, regulatory framework, organizational structure and government processes as a service and bureaucracy, are included in the "business context" step, as part of the imperatives analysis described in section 3.

3 A HIGH-LEVEL FRAMEWORK APPROACH

Analysing imperatives method represents the previous step towards the transformation of an administration body into a SOA-based organization.

We start from Public Administration business context to carry out this process of analysis.

This context shows the organization's business, its functional areas and its organic structure, so as to identify its situation regarding the future government model.

Services supplied by Public Administration to citizens constitute an essential part of this business context. In the same way, we can also observe the determining factors conditioning the organization from both points of view, business and ICT (corporative policy defined at IT level).

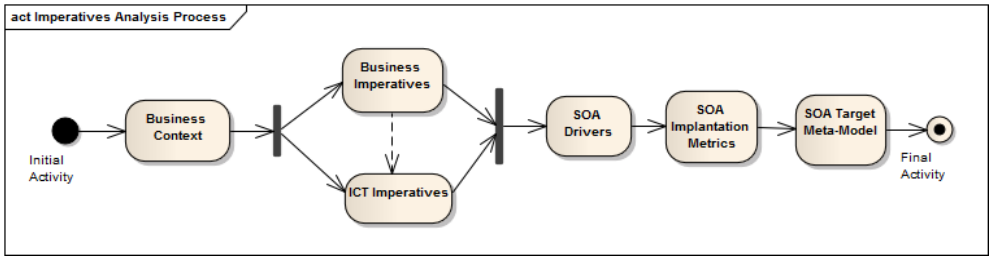


Figure 1: Imperatives Analysis Method.

Once the foundations of the business are laid, we proceed to identify the business imperatives. These refer to the business and service challenges addressed to citizens that Public Administration must face up. We get the ICT imperatives (the translation of the target business into a technical language), after identifying the aforementioned business imperatives, through a deductive process that relates the business challenges to possible initiatives that can be carried out from the ICT area, in order to face them successfully.

These ICT imperatives keep a certain level of abstraction since they define what to do, but not how to do it. In a next step, and, based on ICT imperatives, we will set out the concrete implementations that SOA can do regarding such ICT imperatives. These implementations are called SOA Drivers. This way we can realize the value that a SOA implementation can provide at corporative level, in relation to the need of facing up the established challenges outlined by ICT imperatives, and by extension, by business imperatives.

Therefore, once the relation between SOA Drivers and the own business imperatives has been defined -shown through the different traceability matrices-, we clearly observe that SOA Drivers must act as an engine when implementing SOA. A traceability matrix is a document, usually in a table form, that correlates any two baseline documents that require a many-to-many relationship to determine the completeness of the relationship. It is often applied to high-level requirements and detailed requirements of the product to match parts of high-level design, detailed design, test plan, and test cases. SOA implementation metrics will be defined, after identifying SOA Drivers. The definition and implementation of these metrics are carried out with the aim of evaluating whether the organization is transformed to operate under a SOA paradigm. Each SOA metric represents how efficiency is measured within each SOA Driver. Finally, we will arrive to instantiate the SOA Target Meta-model, once the SOA Drivers cluster has been executed. (Figure 1)

3.1 Business Context

Public Administration, due to its public nature and the service provided to citizens, is defined as *"the set of elements around which the Public Administration defines a global model for the providing of services to citizens, and establishes some strategic targets that determine the shape and minimum levels of providing such services"*.

There is a set of legal and structural elements in Public Administration that will make up the *"core business"* of the administration business context, which must constitute the starting point of our imperative analysis strategy. These common elements to all public administration will be more deeply explained below:

Regulatory Framework. Public Administration is set under a common regulatory framework that not only controls its own operation (internal management), but also the service provided to citizens. Besides, as the public administrative law is especially oriented towards the procedure and it is deeply guaranteed, some of the business procedures of administration bodies do not follow the principles of efficacy and efficiency. Consequently, their re-engineering draws a competitive line when providing these services. For example, in Spain, there are a group of laws that have direct influence on the way services are provided to citizens, such as *Law 11/2007, of 22 June, on electronic access to Public Services for members of the public*, that recognizes the right of citizens to have access to Public Administration by electronic means and regulates the basic points on the use of information technologies within the administrative activity.

All designs associated with services, processes and procedures must be framed within this regulatory background to be implemented, according to the most important concepts these laws point out: service provision, transparency, interoperability and citizens' empowerment.

Charter of Services. They are defined as documents whose purpose is to inform citizens of

the public services. The administration body manages the conditions in which they are provided their rights in relation to these services and the quality commitments offered in terms of its providing. This document is crucial, since it directly frames services that the relevant public administration body supplies and also its real aim as administration, delimiting the business context.

Structure Decrees. Each public administration body is regulated by a structure decree that determines a set of key elements affecting and configuring the SOA Component Diagram. This decree establishes **the competences** to develop the business process (which in turn will become the base of analysis of business imperatives) and the organizational structure that covers these competences (that will determine the Domain Model).

Strategic Plans. They are formal documents approved by the board of government, written to materialize public policies in order to provide citizens with a right. They do not answer to a legal duty, but each administration office should commonly be ruled by one or more strategic or action plans.

The Master Plan for Information Systems. It aims to get a reference framework for the IT development in response to Public Administration's strategic goals.

3.2 Identification of Imperatives

It will be necessary to identify imperatives within the strategy or imperative analysis. It means that we must get imperatives, understood as duty or inexcusable requirements, from the previous business context. Business imperatives can be defined for Public Administration as those that being overcome, may allow reaching excellence in providing citizens with services that inherently correspond to them, as well as in managing internal processes supporting them.

Starting from the analysis of the System Plans and the Competence Plans that Public Administration has been assuming by means of the corresponding decrees of structure and the particular regulatory framework, and, based on the current situation**,** the following business imperatives can be identified in general:

- Fulfilment of operative guidelines that entail measurement, analysis and continuous improvement of the public services offered.
- Checking the current level of administrative procedures by telematics means. It is necessary to perform the foreseen review of telematics procedures that involve the laws
- Existence of not computerized business processes, reengineering and automation of these processes, establishing the degree of agility when computerizing business projects are necessary. It is essential to implement the mechanisms to speed up the business processes computerization in the SOA paradigm.

Once the business imperatives have been established, the next step consists in identifying the ICT imperatives linked to each business imperative. These ICT imperatives are defined as the solutions that the ICT area of the administration body offers to face up the identified business imperatives.

As a result of the business imperatives analysis, and related to them, the next ICT imperatives will be identified, and a traceability matrix will represent the coherence of this identification.

This point is crucial since ICT imperatives will be those that, once translated into the SOA paradigm, will start creating the necessary base for the effective implementation of the architecture in such administration office.

3.3 SOA Drivers

SOA Drivers are initiatives that, under the SOA paradigm, help Public Administration tackle business and ICT imperatives. Therefore, SOA Drivers are the elements that should head the implementation of this architecture and, therefore, carry out the transformation.

A SOA Driver, that is a possible solution associated with each imperative under the SOA paradigm, will be recommended to identify these SOA Drivers for each identified ICT imperative.

To sum up, SOA Drivers will be the tools that aligned with the business context, will provide the organization with different ways of operating, leading to the effective implementation of SOA by instantiating the SOA Target Meta-model.

3.4 SOA Implantation Metrics

SOA implementation metrics (Brow et al, 2008) are the elements used to assess the degree of implementation of the SOA paradigm in the organization. These metrics are defined with the purpose of being able to evaluate the evolution of SOA Drivers in relation to the objective they intend to fulfil. Due to this relation, they are useful to evaluate the SOA implementation state together with its progression along the time.

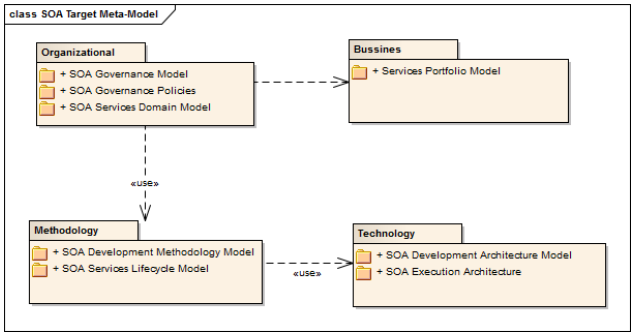


Figure 2: SOA Component Diagram.

Two concepts that will allow us to frame the concept of metrics and their relationship with an organization's KPI and SLA are defined below:

- **KPI** (Key Performance Indicators). It is an element that enables analysing any organization's business processes, making it possible to evaluate the level of achievement regarding the outlined objectives.
- **SLA** (Service Level Agreement). It is an agreement established between the service provider and the consumer by specifying the expected quality of that service.

It is crucial to define a set of implementation metrics that can assess the fulfilment of the identified business goals, in order to carry out a correct implementation of SOA.

The definition of these metrics is planned in the SOA implementation project, included within the activity of defining the SOA Government Model.

The aforementioned definition of metrics must consider the imperative analysis stated in SOA implementation strategy and, particularly, in the SOA Drivers identified.

With regard to the selection of the best metrics and indicators, they must be defined according to, at least, the following criteria:

- Alignment with both, the SOA Drivers identified and the business and ICT imperatives that lie beneath them.
- Possibility of doing homogeneous measurements on metrics, which enable making different comparisons to show how the implementation is evolving.

SOA metrics must be adapted to the main objectives of SOA implementations, so that they will verify the degree of achievement of these objectives.

SOA will define metrics, through its technological infrastructure and its governance policies and procedures, for monitoring services deployed in execution as well as evaluating the degree of implementation of the SOA paradigm into the organization.

The metrics that will allow checking how SOA has been implemented (in a company) can be classified in two different typologies:

- These concerning controlling activities linked to the lifecycle; revision of documentation, additions, versions, detection and use of the existing services, and revision of the accomplishment in the testing phase of the SLA defined.
- These concerning controlling administrative activities and management; tasks related to manage the registry (definition of a service property and maintenance of the registry) and define communication and distribution services policies.

3.5 SOA Target Meta-model (SOA-TM)

We propose a SOA Target Meta-model (Figure 2) for Public Administration that represents the necessary structure of components to start with the operation phase under a SOA paradigm.

This meta-model must be understood both, as part of the definition of the high-level framework and as its own result. That means that it is necessary to define first a set of components that, after executing the imperatives analysis and applying the iterative method proposed, shows the objective final state of the organization, which will operate under a SOA paradigm. This Component Diagram comprises four elements, which refer to the organization (government), business (functional), methodology and technology:

Organizational Components. At this level, policies and procedures that shall structure the SOA government will be defined. They will constitute a set of rules, restrictions and requisites to monitor the behaviour and actions within SOA services, so that

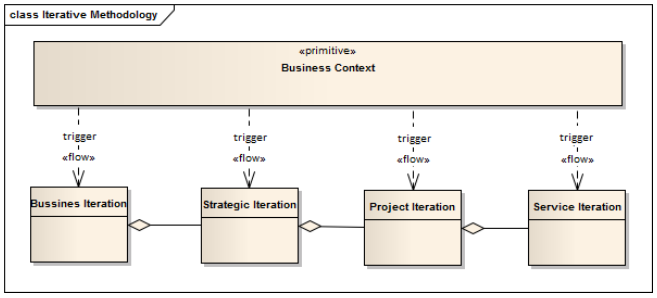


Figure 3: Iterative Strategy.

they can be aligned with the business needs obtaining a successful result after adapting the SOA paradigm. This government model will deal with the definition of SOA Service Domains, based on the legal organic structure of Public Administration. Definition of SOA Service Domain Model.

- Definition of SOA Government Model.
- Definition of SOA Government Policies through the formal definition (Parejo et al, 2011).

Business Components (Functional). This component is based on the analysis of the organization's business processes, obtaining as deliverable:

- Definition of the Services Portfolio.

Methodological Components. This root of the meta-model refers to the need of establishing the methodological bases that allow providing a common development framework.

- Definition of the SOA Services Lifecycle.
- Definition of the development methodology for SOA adaptation, making changes to the lifecycle development. The Ministry of Culture and Sport, for instance, works with methodology focused on the Model-Driven Engineering paradigm, called NDT (Escalona et al, 2008). It covers all aspects regarding the lifecycle development.

Technological Components. At this level, the technological components suitable for the service-oriented development must be selected, giving priority to those that maximize software and platform interoperability reuse.

- Definition of SOA Development Architecture.
- Definition and implantation of SOA Execution Architecture.

4 ITERATIVE STRATEGY

The implementation of a SOA must be considered an incremental process, which cannot be focused on information technologies, but on its own business processes. The strategy for an organization to succeed with this initiative must consist in establishing a time and progression line, where different iterations occur, assuring that technology is aligned with the cultural, organizational and governmental aspects of such organization. Four kinds of iterations must take place (Figure 3) to reach the target of this SOA transformation. These iterations will be part of a lifecycle that will influence the organization in a multidimensional way, will have a periodicity recommended, but not defined and will re-launch the imperatives analysis in different context switches.

This will depend on the business context shooters, which will re-launch the lifecycle according to the level where the event will take place. Each iteration is in turn made up of n-iterations at the next level.

Business Iterations. They constitute a cycle of business analysis that must allow detecting any change of relevance within the organization. From the point of view of PA, the business deals with providing services supplied by executive orders.

Strategic Iterations. They particularly focus on the SOA transformation strategy point defined in the analysis of business and ICT imperatives. This kind of iteration is the result of carrying out multiple initiatives or SOA projects within the framework. As a consequence of this process, technology itself acts as a SOA transformational element. In the same way, it is the opportunity to execute the government policies defined, broaden their acceptance, taking them as organizational and cultural base, and the last targets of each individual project**.**

Project Iterations. The degree of implementation reached on the proposed SOA projects and, therefore, the degree of transformation of the related body will increase in these iterations. They mean a cycle of analysis of the group of SOA projects necessary to fulfil the SOA transformation strategy process of the PA.

Service Iterations. SOA services iterations are implementations of business or technological requirements carried out by the organization, which finally conform the SOA paradigm. This iteration must follow the corporative process defined for the PA in relation to the lifecycle of services as part of SOA methodological definition of development.

5 A PRACTICAL EXPERIENCE

The project presented in this paper was developed by Junta de Andalucía, the regional government of Andalusia, a region located in the south of Spain, counting with more than 200,000 employees. This project was developed by the regional Ministry of Education, Culture and Sport, which is in charge of developing and coordinating public policies both in cultural and sports areas. Among its tasks, this ministry manages public museums, archives, sports clubs, art galleries and theatres. It is also responsible for supporting regional cultural and sports industries, together with other public and private stakeholders. The ICT Department of the Ministry is in charge of running this project, being responsible for all Information Technology policies in the Ministry.

The project started in January 2008 and finished the first complete iteration in June 2009, after instantiating all components of the SOA Target Meta-Model.

Since this moment, the Ministry has executed the iterative methodology in all changes that occurred in the last years, re-launching the lifecycle in relation to the level at which the event takes place, according to the business context shooters.

The results of applying this method to the business context of the Ministry of Culture and Sport shows that the imperative analysis on the e-government services are part of the business context composed by the strategic plan of Andalusia's Culture, the structure decree of Ministry, the Decree 317/2003, November 18th that regulates the Charter of Services and the Open Source decree of Andalusia.

The main conclusion and principal business imperative towards e-government was "*expand the current level of administrative processing by telematics means*". The ICT imperatives associated were "*implementing an administrative procedure engine that allows incorporating of new administrative procedures electronically*". The SOA Drivers defined to fill this deal with "*defining these e-government processes as a set of SOA services that can orchestrate with the aim to compose more complex procedures*".

It is not possible, since it is out of the scope of this paper, to describe all policies, artefacts, processes, roles or actors involved in governance according to the SOA Conformance Office.

The elements of the SOA Target Mata-Model have been created along the process of applying this methodology to all business imperatives. Finally, this model was instantiated. Particularly, 14 domains, 20 processes (governance model and governance policies), a service lifecycle model, a service portfolio definition and new activities for the software development lifecycle, were generated.

From a technological point of view it has changed the software lifecycle based on NDT. A UML element interface type to service modeling was created and activity diagrams corresponding to DAS and DRS were modified. Additionally, a services repository (supported by Enterprise Architect) that is used to identify existing services in the process of analysis was also created.

6 CONCLUSIONS AND FUTURE WORK

This approach will allow us to trace a roadmap towards transforming Public Administration (PA) progressively, beginning with its legal framework and led by the IT, and ending with a model that covers all the required components in SOA.

The results of applying this high-level framework concern the transformation of PA into an organization capable of operating under the SOA paradigm. The ultimate goal of PA regards services provision, especially those related to e-Government and citizen's transparency and empowerment (European Commission, 2010).

Transforming the administration using this method shall ensure that services are provided with the proper level of quality, and that such services are those that should be actually supplied, due to the analysis of imperatives and the correct identification of the business context. Besides, technology will lead the civil servants through SOA Drivers towards the structural change in administration itself in a clear way. This process has proven to be a valuable ally to overcome the human component, integrating technologies and political profiles in the same direction. We have successfully identified the triggers that affect business context and require adapting the organization, through the iterative strategy. The transformation of PA through this

high-level framework will enhance the level of maturity of service provision. Therefore, we can consider levels of maturity in e-government and open government services provision, from the transformation of the organization and the instantiation of this SOA Target Meta-Model.

This will entail, as a future work, the analysis and quantification of the meta-model components that more effectively contribute to reach such level of maturity.

ACKNOWLEDGEMENTS

This research has been supported by the MeGUS of the Ministerio de Ciencia e Innovación and by the NDTQ-Framework project (TIC-5789) of the Junta de Andalucía, Spain.

REFERENCES

Goudos S. K., Loutas N., Peristeras V., Tarabanis K., 2007. Public Administration Domain Ontology for a Semantic Web Services E-Government Framework Services Computing. *SCC 2007. IEEE International Conference on, vol., no., pp.270-277*.

Popescu D., Nirvana P., Ciprian D., 2012. E-Frameworks to Optimize Public Administration Services. *Digital Democracy: Concepts, Methodologies, Tools and Applications (3 Vols.). IGI Global.*

Castellano M., Pastore N., Arcieri F., Summo V., de Grecis G. B., 2005. An E-Government Cooperative Framework for Government Agencies, System Sciences. *HICSS '05. Proceedings of the 38th Annual Hawaii International Conference vol., no., pp. 121*.

Arsanjani A., 2004. *Service-Oriented Modeling and Architecture*. http://www.ibm.com/developerworks/webservices/library/ws-soa-design1/

Marks E. A., 2008. *Service-Oriented Architecture Governance for the Services Driven Enterprise*. John Wiley & Sons.

Sedeño J., Torrecilla-Salinas C. J., Escalona M. J, Mejías M., 2013. An approach to transform Public Administration into SOA-based organizations. *JISBD2013. Proceedings of the 18th Conferences of Software Engineering and Data Bases.* Madrid.

Brow W. A, Laird R., Gee C., Mitra Y., 2008. *Governance: Achieving and Sustaining Business and IT Agility*. IBM Press.

Parejo J. A., Fernández P., Ruiz-Cortés A., 2011. WS-Governance: a policy language for SOA Governance. *ICSOC2011. Proceedings of the 9th Int. Conference Service Oriented Computing*. Springer-Verlag.

Escalona, M. J., Aragón, G., 2008. NDT: A Model-Driven Approach for Web Requirements. *IEEE Transactions on Software Engineering. Vol. 34. No 3. pp. 370-390*.

European Commission, 2010. The European e-Government Action Plan 2011-2015 Harnessing ICT to promote smart, sustainable & innovative Government.

Hall J. L., 2007. Implications of success and persistence for public sector performance. *Public Organization Review, 7(3), 281-297.*

Tregear R., Jenkins T., 2007. Government process management.http://www.w.bptrends.com/publicationfiles/10-07-ART-Govt.ProcessMgt. -Tregear and Jenkins-ph.pdf.

Repa V., 2006. Modeling business processes in public administration. *Advances in information systems development (pp. 107-118)*. Berlin. Springer.

Galster M., Laurens L., Paris A., 2013. Service-oriented architecture in variability-intensive environments: pitfalls and best practices in the example of local e-government. *IEEE Software. IEEE computer Society Digital Library.* IEEE Computer Society.

Alwadain, A., Fielt, E., Korthaus, A., & Rosemann, M., 2013. A Comparative Analysis of the Integration of SOA Elements in Widely-Used Enterprise Architecture Frameworks. *International Journal of Intelligent Information Technologies (IJIIT), 9*(2), 54-70. doi:10.4018/jiit.2013040105.

Abdul-Manan M., Hyland P., 2013. A Framework for Assessing Enterprise-Wide SOA Implementation Readiness. *International Journal of Intelligent Information Technologies (IJIIT), 9*(2), 21-37. doi:10.4018/jiit.2013040103.

Koumaditis K., Marinos T., Rupino Da Cunha P., 2013. SOA implementation critical success factors in healthcare. *Journal of Enterprise Information Management, Vol. 26 Iss: 4, pp.343 - 362.*

Ruas de Oliveira L. B., Romero K., Feitosa D., Yumi E., 2010. Reference Models and Reference Architectures Based on Service-Oriented Architecture: A Systematic Review. *ECSA2010. In proceeding of: Software Architecture, 4th European Conference, Copenhagen, Denmark*. Proceedings Source: DBLP.

Delgado H., Losavio F., Matteo A., 2013. Goal oriented techniques and methods: Goal refinement and levels of abstraction. *Computing Conference (CLEI), XXXIX Latin American , vol., no., pp.1,12, 7-1.*

Alan McSweeney, 2000. Enterprise Architecture and TOGAF. *The Open Group Architecture Framework.*

A Flexible System for a Comprehensive Analysis of Bibliographical Data

Sahar Vahdati, Andreas Behrend, Gereon Schüller and Rainer Manthey
Institute of Computer Science III, University of Bonn, Römerstr. 164, D-53117 Bonn, Germany
{*vahdati, behrend, i3schuel, manthey*}*@cs.uni-bonn.de*

Keywords: Bibliographic Database, Digital Library, Citation Analysis.

Abstract: Scientific literature has become easily accessible by now but a comprehensive analysis of the contents and interrelationships between research papers is often missing. Therefore, a time consuming bibliographical analysis is usually performed by scientists before they can really start their research. This manual process includes the identification of the most important research trends, major papers, auspicious approaches, established conference series as well as the search for most active groups for a specific research topic. In addition, scientists have to collect related academic literature for avoiding reinvention of already published results. Although a large number of literature management systems have been developed in order to support researchers in these tasks, the offered analysis of bibliographical data is still quite limited. In this paper, we identify some of the missing analysis features and show how they could be implemented using data about author affiliations, reference relations and additional metadata, automatically generated from a set of research articles. The resulting prototypical implementation indicates the way towards the design of a general and extendible bibliographic analysis system.

1 INTRODUCTION

Bibliographic analysis is a very important aspect in scientific work but usually a time-consuming process. There are a lot of scientific literature management systems available by now supporting researchers in this process to a certain extent (e.g, ACM Digital Library, DBLP, CiteSeerX, SpringerLink, Google Scholar or Microsoft Academic Search, but a comprehensive analysis of bibliographical data is still missing. Most of the systems focus on a perfect keyword search or the computation of certain impact factors which do not help researchers in getting a good overview about a certain research field (Bakkalbasi et al., 2006). In contrast, researchers are often interested in getting a quick overview about a scientific topic rather than a list of papers ordered according to their citation ranks.

In particular researchers would like to find out: "What are the most active conferences, groups, institutes in the research area of my interest?" "Who was the most influential author and who is it now?" "What papers are the most important ones for a given research topic?" "Are there different schools or research directions and which direction has turned out to be a dead end, retrospectively?" "Which research topic is currently en vogue establishing a kind of trend?" "What conferences should I choose if I want to submit a paper about a certain topic?" . The basic technique needed to get such information is the exploration of related work which starts with navigating the citations provided in a set of research articles. To this end, bibliographical databases, digital libraries and search engines could be used which offer a citation analysis of publications (Lister and Box, 2008). Such systems offer different search methods but usually leave researchers alone with long lists of papers possibly matching the provided search criteria. For example, Google Scholar is good in finding papers which are relevant to a given list of keywords. In addition, the number of citations by other papers are displayed for each document which leads to another list of papers.

Despite of these fine-tuned results, the questions stated above cannot be answered that way and the user needs to perform various similar searches before he gets an overview about the research field he is interested in. Microsoft offers another search engine which additionally provides a graphical visualization for exploring the citation dependencies stepwise in one direction. The system, however, cannot provide a general overview over certain research areas and allows for investigating the interrelations of one selected paper, only. In this paper, we show how missing analysis features could be implemented using data about author affiliations, reference relations and additional metadata automatically generated from re-

search articles. All paper related data is stored in a relational database and SQL queries are employed allowing for a flexible and comprehensive article analysis. In contrast to other literature services, our approach supports

- the analysis of the complete citation graph allowing for detecting research schools, dead-end research directions or methodological differences
- the distinction between different types of citations which allows for a better understanding of relationships among papers
- the automatic detection of key papers as well as key conferences respectively journals for a given research school
- the analysis of author-related information (e.g. affiliations).

At the end of this paper, we provide an evaluation of our prototypical implementation using a sample data set which contains documents from the research field deductive databases over the last 37 years. We believe that our system indicates the way towards the design of a general bibliographic analysis tool.

2 ARCHITECTURE

The architecture of our system consists of three main components: Base Data Collection, Bibliographical Database System and User Interface. The first level is responsible for collecting, editing, storing and indexing research papers. To this end, paper-related information as well as a digital copy of each paper is stored in the repository after passing a digitization process. Data acquisition is one of the crucial steps as it determines the quality of the analysis results later. We have employed metadata extraction tools to our sample document set in order to obtain detailed information about title, authors, affiliation, publisher, references and covered research topics. Additional metadata could be extracted from other external systems such as digital libraries (DL), bibliographical catalogues or literature databases. In fact, we plan to develop our system towards a metasearch engine which integrates the analysis features of other DL. In order to enhance the paper classification process, we have annotated papers by information about the paper contents such as keywords as well as data model or programming language used in the presentation. Most of this information has been automatically extracted but manually checked in order to have a very clean data set. This additional data allowed for further refining the search for relevant articles with respect to a given

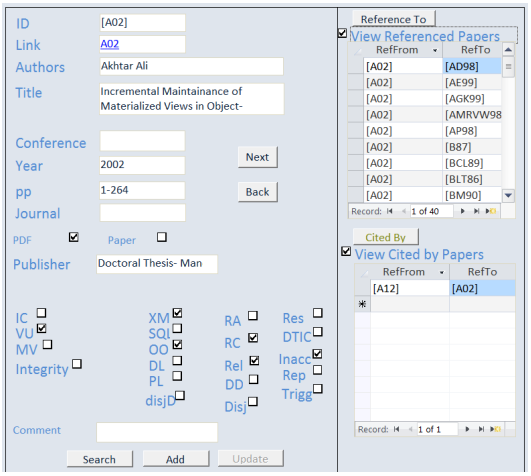


Figure 1: The interface of our prototype for advanced content search.

topic. Finally, the user interface provided is able to visualize the results using different types of graphs such as intensity maps, line or network graphs which transfer the analysis to a human readable format.

3 SAMPLE DATA COLLECTION

We have used a specific sample set of papers for showing the expressiveness of our bibliographic analysis tool. To this end, we used a collection of papers which are concerned with three classical deductive database topics, namely materialized view maintenance, integrity checking and view updating.

3.1 Conceptual Design

Although the analysis of paper-related information such as citations and topics covered were the main focus of our work, we additionally wanted to support the detection of the most influential conferences and journals for particular research fields. Therefore, we needed to carefully distinguish between the various forms of scientific publications such as proceedings, books or journals. Furthermore, we had to extract some additional metadata from the papers such as citations, keywords or author affiliations which are usually not provided by similar systems for scientific literature management. Metadata extraction can be automatically done using extraction tools such as TeamBeam. In our sample data set we could finally use more than 30 attributes for characterizing a paper including authorship, affiliations, citations, keywords and potentially covered research topics. Aside from the main paper table, further tables are employed

for storing the metadata determined such as reference lists or the information about authors (in order to handle multiple authorship). Based on that data set, we could provide a much deeper paper analysis using SQL queries than offered in similar bibliographic systems available today. Our data collections contains 1103 papers published from 1975 till 2012. The reference list basically forms the edge relation of the underlying citation graph and additionally contains a Boolean attribute indicating the "importance" of the citation relationship.

In fact, citations may play a different role in a paper. Some of them are used for indicating that the authors know other research work (or traditions) which is related to their own paper. Other citations are used to refer to scientific material (methods, approaches) which is really necessary in order to understand the presented approach of the paper under investigation. We consider the latter as much more important as they indicate research directions and even real scientific cooperations. During the process of evaluating the reference type we used the value "more important" as default which is automatically replaced if the citation appears only once and the paper keywords are considerably different. More details about the automatic detection of reference types is given in Section 4.2. Another important category of citations are self-references which are consequently omitted within the entire analysis process.

Many of the popular online services suffer from unreliable data. In order to avoid similar problems and to provide more reliable analysis results, we have applied a data cleaning and correction step (a step towards semi-automated curation). To this end, the stored information about publishers, page numbers and authorship has been verified using different online sources. Another problem we had to solve before analyzing the data was the generation of unique paper IDs. The primary ID of each paper is automatically generated from the authors' names and the publication year. In case that this method would lead to duplicates, the naming method is stepwise refined following the recommendations from (Han and et al., 2004) until all conflicts have been resolved. Although this represents no general solution, we could already obtain a clean sample set this way.

3.2 Content Description

In the following two sections we show how complex bibliographical questions can already be answered using the automatically collected paper data from above. For example, in Section 4 we show how the citation graph could be used to identify different research schools for a particular research topic. This kind of paper analysis, however, could be even improved if more information about the paper's contents would be present. We used additional classification data for each paper in order to provide better query results to the user. To this end, we employed the classification data from the digital library of ACM.

In order to provide even more classification details we determined characteristic keywords within the paper related to the employed data models, programming languages, and algorithms. This classification process was done using a pre-defined set of keywords (e.g. Relational Algebra (RA), Update Propagation (UP), Fixpoint (FP), SLDNF, Deductive Rules (DL), Transformation-based (TB),...) and text-mining tools provided by the underlying database system. The resulting data provide a more precise content description such that researchers can perform more advanced paper searches while reducing the number of false positives in the list of returned answers.

For example, in our user interface you can search for papers concerning integrity checking where the authors use the relational data model as well as the relational algebra in their presentation (see Figure 1). This way, the article "[Bro00]" with the title 'A general treatment of dynamic integrity constraints' and the article "[Dec02]" with the title 'Translating Advanced Integrity Checking Technology to SQL' could be easily identified as very related papers despite of the very different values in the ACM classification system (G.2.0 and H.2.1 vs. H.2.7 and H.2.3).

4 CITATION ANALYSIS

Despite of the fact that many scientific literature management tools provide a citation number for each paper, a comprehensive citation analysis is usually missing. Having the entire citation graph at hand, various interesting conclusions about the importance of paper ideas, authors, research schools and the success of research directions can be drawn.

4.1 Building a Citation Graph

Reference Depths: Citations form a relation between papers leading to a directed dependency graph. Papers may be connected via several relationships (paths) with different distance values. We call a relationship between papers "non-direct references" if a paper does not refer to a paper in its own reference list but indirectly via one or several referenced papers (i.e., the distance is larger than 1).

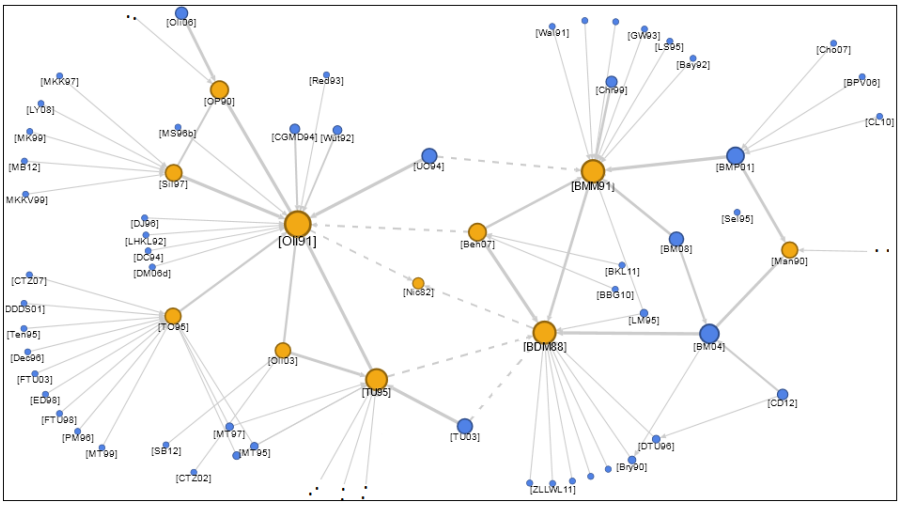


Figure 2: Citation relationships in form of a network graph view (Google Fusion Tables).

Self-References: The self-references may be justified if an author refines his previous work. However, self-references are often considered to be not that valuable. We developed a filter query for avoiding the analysis of such references and stored the remaining references using the materialized view "NewReferences".

Ancestors: All direct references which have been extracted from the documents have been added to the References table with distance number 1. These references can be used to determine indirect relations between papers that can be modelled as ancestors and descendants with different distances. For the subsequent analysis queries, it would be desirable to have the ancestors and descendants relation in materialized form. The paper relationship graph, however, is highly connected and a paper may be related to another paper by various paths of different lengths. Therefore, we created a new table named "NewAncestors" which contained all paths with the smallest distance. The corresponding paths have been determined using a fixpoint iteration starting from all edges in table "NewReferences" with distance 1.

The respective query is iteratively applied as long as there are new insertions. With our test data, the iteration loop ended after 8 iterations, which means that the maximal distance between two related papers was 8. The total number of direct relations between two papers in our set is 4552, while there are more than 35000 indirect relations. These relations between papers form a graph in which papers are the nodes and the distance is the weight of each edge between two related papers. A small fraction of our reference relation can be seen in Figure 2. The yellow nodes indicate that this paper is referenced by another one. The more papers refer to a yellow node, the bigger (and more important) the node becomes. If a paper is not referenced at all, the node gets the color blue.

4.2 Analyzing Citations

The stored direct references together with the computed indirect ones allow for standard computations such as "Who is the most referenced author for a particular topic?" or "What paper about integrity checking has the highest number of references?". Figure 3 shows the list of papers with the highest number of citations, including the authors, conferences, and publication years. Here we have already used the total number of references including the computed indirect ones. Our data could also be used to find the most active author who is still publishing in the respective research fields. For example, the most active researcher over the last two years 2011 and 2012 in the field of integrity checking was Hendrik Decker according to the data from our collection.

The citation graph, however, even allows for a deeper paper analysis. Suppose you want to know whether there are different schools (or approaches) for checking integrity violations efficiently. In Figure 2, clearly two subgraphs can be identified which basically represent two research schools for this particular research area. As mentioned above, we have used two types of references in order to consider the different importance for the respective paper. Solid lines are used to indicate important citations (strong relationships) which are necessary to understand the presented approach while dotted lines (weak relationships) are employed for citations of minor importance. The paper "[TU95]", therefore, belongs to the subgraph centered by "[Oli91]" and not to the other one with the center "[BDM88]" because of the dot-

RefTo	NumRef	Authors	Conference	Journal	Year
[Nic82]	129	Nicolas		AI	1982
[BLT86]	116	Blakeley, Larson	SIGMOD		1986
[GMS93]	103	Gupta, Mumick,	SIGMOD		1993
[BS81a]	97	Bancilhon, Spyr		TODS	1981
[ZGHW95]	88	Zhuge, Garcia, H	SIGMOD		1995
[GM95]	87	Gupta, Mumick		DE	1995
[DB82a]	81	Dayal, Bernstein		TODS	1982
[GL95]	78	Griffin, Libkin	SIGMOD		1995

Figure 3: Papers sorted according to the number of references.

ted connection to the latter. For determining the two connected components of the depicted citation graph, current database systems already provide methods (e.g., in Oracle 11gR1 as part of the *Oracle Spatial and Graph* facility) which have to be applied to strong relationships, only.

In order to distinguish weak and strong relationships, the references have been weighted using a bonus system. To this end, we employed the following formula $bonus(c,k,o) := 10*c+20*k+20*o$, where c is the total number of common coauthors, k is the number of common keywords and o is the number of reference occurrences within the referencing paper. The first subterm is based on the assumption that common coauthors indicate related paper contents. The factor 10, however, leads to a relatively small impact on the total bonus number because coauthorship is only a very rough indicator. The main criterion is the number of common keywords used in the two papers as it provides the best indicator for the paper topic. We employed again a list of pre-defined keywords like UP, FP, SLDNF, RA, Deductive Rules and Transformation-based (cf. Section 3.2) each of them contributing to the total bonus by the weighting factor 20.

This rough estimation worked well for our test scenario but could be further improved by more sophisticated IR approaches (like the tf-idf method (Wu and et al., 2008)) for determining the importance of a keyword within a document. The third subterm evaluates the number of occurrences of a certain reference. If authors are referenced several times, we may assume that the relationship to this previous work is high. We simply added all bonus points and indicated a reference relation as weak if the resulting weight is below 300. In addition, weights above 700 are indicated by even thicker solid lines in order to show very close relationships between two papers. On that basis, different research schools could be easily identified as shown in Figure 2. The used interval [300,700] worked almost perfectly for our sample set but an extended evaluation is still to be done. This way, not only different research schools could be identified but also the importance of certain approaches may be estimated. For example, the subgraph on the left-hand side contains considerably more nodes than the one on the right-hand side, making it potentially more influential (active) than the other one. Furthermore, we could use our filter to omit any self-references which would allow to estimate the number of people in the respective research group working on that particular subject. Despite of these already promising results, the approach can be certainly further refined by applying more sophisticated methods for estimating paper relationship such as the vector space model from IR.

5 RESULTS

In the following, we provide a brief overview of our analysis results. While the following presentation remains rather sketchy, it already shows the general possibilities of a comprehensive paper analysis.

5.1 Publishers and Venues

The documents in our collection had been published in 314 different conferences and journals. There are 199 conferences related to the three research areas and five top active conferences could be determined according to the number of references to papers published there. The conferences in which most of the papers were published over the years are VLDB, SIGMOD, ICDE, PODS, and DEXA. In particular, around 1/5 of all papers from our collection appeared in these conferences.

In Figure 4 the line graph of the number of papers published in the top 5 conferences is shown. During the mid 90s there was a peak in the interest for the respective three topics. Then, between 2003 and 2007 there was a renewed interest in these topics with publications at ICDE, VLDB and DEXA.

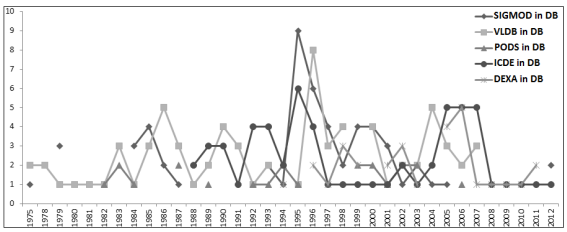


Figure 4: Number of papers from our collection published in the top 5 conferences.

Afterwards, the number of published papers at these conference decreased significantly, showing that the research topics under investigation have lost their dynamics. This analysis is equivalent to the observations made by (Mayol and Teniente, 1999). During the last two years 2011 and 2012, the confer-

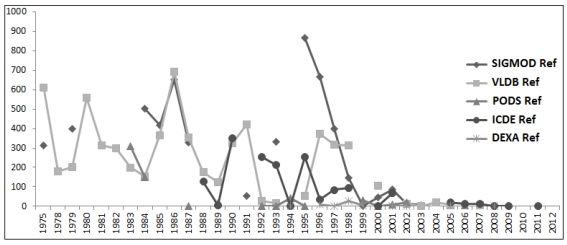


Figure 5: No. of papers referencing conferences per year.

ence series KES (Conference on Knowledge-Based and Intelligent Information & Engineering Systems) became more important for publishing results from our three sample topics. So, if a researcher is interested in the topics integrity checking, materialized view maintenance or view updating, he or she ought to consult older papers from the identified top 5 conferences in order to get an overview about the proposed research approaches. For publishing new results, the conferences KES, DEXA and ICDE appear to be more appropriate than VLDB or PODS.

5.2 Active Conferences

Let us now consider the effectiveness role of conferences for the topics under investigation. To this end, we considered the top 10 conferences and determined how often papers in these conferences have been referenced by papers from our collection. Two queries are employed over the table Papers and NewAncestors which comprise direct as well as indirect references while avoiding the counting of self-references. The first query returns the number of references from different papers to papers published by these conferences and journals. The second one illustrates the number of papers published by a certain conference with respect to the year of publication. The number of references to different conferences is counted in the first query which performs an inner join of NewAncestors on the table Paper. Afterwards the result is aggregated in order to sum up the number of references for a particular conference with respect to each year. The result was that the most active conferences are also the most referenced ones, namely VLDB, SIGMOD, ICDE, PODS, and DEXA.

In Figure 5 the number of references to the 5 most active conferences per year is shown. The oldest and still active conference in our interest research area is VLDB. For example, papers from the VLDB 1990 were cited by 100 different papers from our collection. The conferences which have been mostly referenced by papers from our collection are SIGMOD and VLDB indicating that most of the underlying methodological framework was presented at these conferences. Another interesting aspect is the aging of citations. More recent conferences receive very few citations indicating that new publications on our preselected topics do not use results published in these conferences anymore but rather refer to classical standard papers from the past. Note that there are other ways of analyzing co-authorship which have been proposed, e.g., (Nascimento et al., 2003; Smeaton and et al., 2002), which could be integrated in our system, too.

5.3 Active Countries

Generally, a paper arises from a work or project which has been done in a research institute or university. The origin of a paper is usually indicated in its metadata. The information about the organizations in which the papers have been produced allows for determining the amount of research activity of that organization or country in a certain way. Having such information stored in the database enables us to identify the most active countries for a particular research area.

Another interesting result is the identification of topic movements over countries. By publishing the original papers in a conference or journal, other researchers become aware about that new topic and start working on that. Consequently, a topic may become a trend in a country's research for a particular time interval which is reflected by the intensity of the respective publication numbers.

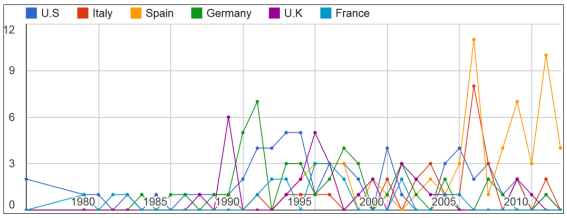


Figure 6: Topic movement for integrity checking over time and country

Figure 6 illustrates the top 6 active countries which have published the highest number of papers over years. In 1975, the first paper on integrity checking appeared from the USA. In parallel, researchers in France were active in this research field and published papers in the same year. Afterwards, other countries like Germany and Spain became interested in this topic and are even still working on them. Another interesting point is that Spain represents the most active country nowadays for the field of Integrity Checking, although it was the last country in the list starting to publish paper in this research area. It is also visible that this topic was an active research field during the 90s but it regained interest recently.

5.4 Citation Genealogy

The impact of scientific publications (Aksnes, 2005) is often estimated by the number of direct citations they receive. On the other hand, indirect references also indicate the relationships of two papers. Suppose a user wants to know whether a paper indirectly refers to a very important work in a certain research field. A possible way to answer this question could be to determine all indirectly referenced papers and to select the one with the highest impact.

In fact, public bibliographic services such as the academic research system by Microsoft allow for exploring indirect citation references by hand. However, this may soon turn into a very tedious task and a query about the most influential paper which is indirectly cited can hardly be answered that way. Indirect citation references may also show the footprints of a publication topic through the evolution of its respective research area. The same technique can be used for analyzing the cited_by relationship. Suppose you want to know what is the most important paper (directly or indirectly) citing the publication under investigation. Again, a query like this could be easily implemented using the ancestor relationship among paper citations

```
SELECT NA.RefFrom, NA.RefTo, max(NA.#DC)
FROM   NewAncestors AS NA
GROUP BY NA.RefTo;
```

where the attribute #DC refers to the number of direct citations. This type of query is currently not supported by any of the public bibliographic services available so far. In fact, most of the systems just focus on direct relationships neglecting the influence of transitive connections. The exploration of indirect references could be used to refine the impact values of a paper but it is not sufficient just to consider the total number of direct and indirect ones. The reason is that once a paper is cited by a key paper with many direct references, this paper inherits all this references which may lead to unjustified high impact scores.

In Fig. 7 the ranking of papers on the topic 'materialized view maintenance' is depicted according to their total number of direct as well as indirect references. We have highlighted the papers we consider to be key research papers in the respective research area. Obviously, both measurements provide valuable information about possible key papers. The number of indirect references in particular indicates how active a research direction has been followed and whether the ideas of a highly cited paper have really spread. The computation of a general impact factor, however, is a difficult task and various other paper related values should be additionally considered. At least, in our

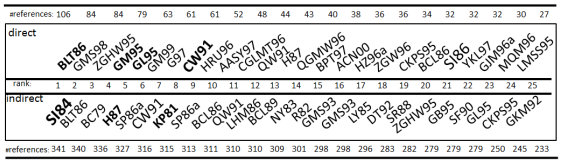


Figure 7: Citation Genealogy for Selected Papers.

particular case Data mining techniques could be used in order to discover a formula for computing impact factors based on direct and indirect references (Chen, 2006). The resulting formula should also take an aging factor into account. For example, the length of the time interval in which direct references occurred may indicate the relevance of a paper and its addressed research topic.

The ancestor relationship could also be used to form a kind of scientific genealogy for any tracked publication. In this way one can see all the work that directly or indirectly influenced a given paper. In Figure 8 we present a sample case in which the ancestor relationship for four different papers is depicted. To this end, we have chosen four papers with a very high number of (in)direct citations (cited_by as well as cited_to references). In the sample study, we have chosen four papers with more than 100 references each, as this border allowed to determine almost every key research paper in our sample collection. It is interesting to notice that the ancestors and children for the rightmost key paper (root) showed almost no interconnections with ancestors and children from another root paper.

This disconnection indicates two very different research fields (integrity checking for the first three papers and query evaluation for the one on the right) as the authors did not refer to the other branch. Thus, different research schools or research fields could be identified even though no keywords nor specific terms were known in advance. The resulting tool could be used to identify unknown research schools and to automatically determine a representative collection of keywords for them (just like reverse engineering). We have called the respective graph *Citation Genealogy*. It has been prepared using the TouchGraph[1] software. In principle, four different cases may occur depending on possible overlaps between the ancestors and/or the children subgraphs. In case that we have connections between children and ancestors of two different key papers (see the two key papers from left-to-right in Fig. 8) we may assume that we are dealing with just one research field from which more than one key paper originated. The case that we have no connections between ancestors nor children is a strong indication

[1] http://www.touchgraph.com/navigator

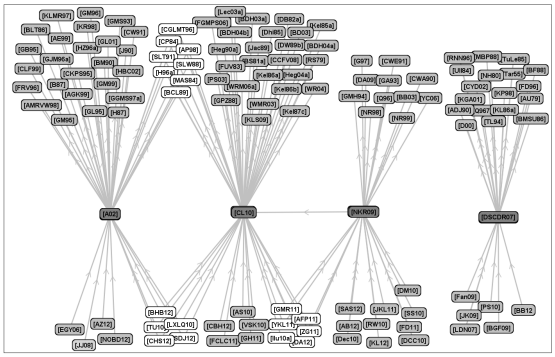


Figure 8: Citation Genealogy for Selected Papers.

for two different research schools. Another case is that we have some connections between the children and even the key papers but none between the ancestor papers (as it is the case for the two key papers in the middle of Fig. 8). This constellation may indicate that two different research directions have developed over time from one common research field. Note that this kind of analysis could even be refined by omitting weak reference types as proposed in Section 4. In this way, however, already identified keywords for different research topics are implicitly used again.

6 RELATED WORK

For decades, citation counts and impact factor scores ,e.g., the h-index or Eigenfactor have been the primary currency in the entire business of publication (Rahm, 2005). Of course, the corresponding metrics could be easily incorporated into our analysis queries. The problem of evaluating the value of paper references has been tackled in various publications ,e.g., (Lacasta and et al., 2013) and (Newman, 2001). In particular, these authors try to find criteria which help to discover very similar scientific work or even duplicated results. An approach for automatically extracting topic classifiers from a paper text has been proposed in (Klampfl, 2013) that could be used in our system for refining the distinction of weak and strong references. The problem of reference reconciliation and methodes for detecting similar references and improving keyword search has been proposed in (Dong and et al., 2005), (Tejada and et al., 2002) and (Harzing, 2013) which allow to incorporate similarity measures. In (Falagas and et al., 2008) and (Jacso, 2005) comparisons between four popular online bibliographical databases has been done using specific keyword searches and analyzing the utility of the retrieved information. The problem of search engines returning lists of results which needs additional manual mining is also addressed with an enriching approach in (Khazaei, 2012). All these approaches, however, solve particular search problems only. A system that provides a comprehensive and extensible citation analysis, however, is still missing. We decided against the usage of a graph database because of the extensive multi-user support we need in our system and the lack of support for large scale application in graph databases (Vicknair and et al., 2010).

7 EVALUATION

For evaluating our system two aspects have to considered: performance and accuracy of the results. With respect to performance, the only critical part is the determination of the transitive closure of the reference relationship. As the resulting graph is materialized, all queries using this information (see samples in Subsections 5.3 and 5.4) could be executed in less than a second using a standard PC. This is not surprising because of the relatively small amount of around 35.000 paths for the 1103 papers and applicable index structures which would scale well for even bigger collections such as DBLP. The determination and materialization of the transitive closure took less than 5 Minutes by exploring link connections of at most 8 steps. Another data set from DBLP has been tested which contained more than 2300 papers with only 6000 [2] indirect references and detected 8 again as the maximal distance of two related papers. Even if the maximal distance number were considerably higher in different research fields, the computation of the transitive closure remains to be strongly limited and we can assume that the total number of indirect relations may considerably stay below $|collection| \times |collection|$.

Furthermore, the reference list of already processed papers hardly changes anymore such that the transitive closure of them really has to be computed only once. New paths within the transitive closure induced by new papers added to the collection can be efficiently determined using incremental maintenance methods as proposed, e.g., in (Behrend, 2011). In addition, the determination of the transitive closure could be the application of recursive views as supported by many database systems like PostgreSQL by now. In contrast to performance issues, the evaluation of the system's accuracy is much more problematic due to the specific data collection we currently work on. Having preselected papers from the research

[2]The low number is explained by the fact that this was a collection with randomly chosen papers. Thus, the number of related papers was quite small in contrast to our topic specific collection.

fields materialized view maintenance, integrity checking and view updating, we can hardly evaluate the accuracy of finding other research fields. For this we would need to evaluate a broader collection of papers. The detection of research schools, however, was highly accurate for our collection but an evaluation of our rules for distinguishing weak and strong reference types is still to be done. In particular, the trade-off between the detection of as many different schools as possible while avoiding the generation of false positives has still to be investigated. Nevertheless, the results we have achieved so far already indicate that the automatic detection of research schools and/or research fields is possible in a feasible way.

8 CONCLUSION

We have presented a system for bibliographical analysis which is desired by many researchers. Our analysis results indicate that such a system is able to answer the bibliographical questions which researchers might encounter while searching for specific papers in a particular area. Our tool can become a powerful and influential information system supporting researchers in other scientific communities to collect, manage, access the bibliographical and citation analysis of documents. The specificity of our system proposal is the consequent application of a database system with a full-fledged query language at hand. In this way, a flexible analysis tool could be programmed which can be easily extended by new paper related measurements. While the analysis part as well as data acquisition is constantly extended, we additionally plan to develop a web-based interface for supporting a flexible querying in the future.

REFERENCES

Aksnes, W. (2005). *Citation and Their Use as Indicators in Science Policy. Study of Validity and Applicability Issues with a Particular Focus on Highly Cited Papers*. PhD thesis, University of Twente.

Bakkalbasi, N., Bauer, K., Glover, J., and Wang, L. (2006). Three options for citation tracking: Google scholar, scopus and web of science. *BDL*, 3(1).

Behrend (2011). A uniform fixpoint approach to the implementation of inference methods for deductive databases. In *LNAI*, pages 1–16.

Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, 57(3):359–377.

Dong and et al. (2005). Reference reconciliation in complex information spaces. *SIGMOD Rec.*, pages 85–96.

Falagas and et al. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *FASEB.*, 22(2):338–342.

Han, H. and et al. (2004). Two supervised learning approaches for name disambiguation in author citations. In *JCDL*, pages 296–305.

Harzing (2013). A preliminary test of google scholar as a source for citation data: a longitudinal study of nobel prize winners. *Scientometrics.*, 94(3):1057–1075.

Jacso (2005). As we may search-comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases. *CURRENT SCIENCE-BANGALORE.*, 89(9):1537–1547.

Khazaei, H. (2012). Metadata visualization of scholarly search results: supporting exploration and discovery. In *i-KNOW*, pages 1–8.

Klampfl, K. (2013). An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. In *TPDL*, pages 144–155.

Lacasta and et al. (2013). Design and evaluation of a semantic enrichment process for bibliographic databases. *DKE*, 88(1):94–107.

Lister, R. and Box, I. (2008). A citation analysis of the sigcse 2007 proceedings. In *SIGCSE*, pages 476–480.

Mayol, E. and Teniente, E. (1999). A survey of current methods for integrity constraint maintenance and view updating. In *ER Workshops*, pages 62–73.

Nascimento, M. A., Sander, J., and Pound, J. (2003). Analysis of sigmod's co-authorship graph. *SIGMOD Rec.*, 32(3):8–10.

Newman (2001). The structure of scientific collaboration networks. In *PNAS*, pages 401–409.

Rahm, T. (2005). Citation analysis of database publications. *SIGMOD Rec.*, pages 48–53.

Smeaton, A. F. and et al. (2002). Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36:39–43.

Tejada and et al. (2002). Learning domain-independent string transformation weights for high accuracy object identification. *SIGMOD Rec.*, pages 350–359.

Vicknair and et al. (2010). A comparison of a graph database and a relational database: a data provenance perspective. *ACMSE.*, pages 1–6.

Wu and et al. (2008). Interpreting tf-idf term weights as making relevance decisions. 26(3):1–37.

Implementing a Semantic Catalogue of Geospatial Data

Helbert Arenas, Benjamin Harbelot and Christophe Cruz
Laboratoire Le2i, UMR-6302 CNRS,Département d'Informatique , Université de Bourgogne,
7 Boulevard Docteur Petitjean, 21078 Dijon, France
{*helbert.arenas, benjamin.harbelot*}*@checksem.fr, christophe.cruz@u-bourgogne.fr*

Keywords: CSW, OGC, Triplestore, Metadata.

Abstract: Complex spatial analysis requires the combination of heterogeneous datasets. However the identification of a dataset of interest is not a trivial task. Users need to review metadata records in order to select the most suitable datasets. We propose the implementation of a system for metadata management based on semantic web technologies. Our implementation helps the user with the selection task. In this paper, we present a CSW that uses a triplestore as its metadata repository. We implement a translator between Filter Encoding and SPARQL/GeoSPARQL in order to comply to basic OGC standards. Our results are promising however, this is a novel field with room for improvement.

1 INTRODUCTION

There is a growing interest in the development of the SDI (Spatial Data Infrastructure), a term that refers to the sharing of information and resources between different institutions.The term was first used by the United States National Research Council in 1993. It refers to the set of technologies, policies and agreements designed to allow the communication between spatial data providers and users (ESRI, 2010).

Currently vast amounts of information are being deployed in the internet through web services. However, in order to profit of this information, potential users need to first identify relevant and suitable datasets. Later, researchers and decision makers would be able to implement *smart queries*. This is a term first employed by Goodwin (2005). It refers to the combination of heterogeneous data sources in order to solve complex problems (Goodwin, 2005).

In the spatial domain this has been possible, thanks, in a significant part to the standardization efforts by OGC (Open Geospatial Consortium). OGC is an international industry and academic group whose goal is to develop open standards that enable communication between heterogeneous systems (OGC, 2012). The tasks in which OGC is interested are: publishing, finding and binding spatial information. OGC provides standards that allow data providers and users to communicate using a common language. The information is offered through web services such as WFS (Web Feature Service), WMS (Web Map Service) or SOS (Sensor Observation Service). However, in order to identify a dataset of interest the user needs first to identify it, using a catalog service. The OGC standard for catalog services is CSW (Catalogue Service for the Web).

OGC defines the interfaces and operations to query metadata records. There are both commercial and opensource/freeware CSW implementations. Among the commercials we can find ESRI ArcGIS server and MapInfo Manager. Among the opensource implementations we find Constellation, Degree and GeoNetworkCSW. The OGC standards do not indicate specific software components. In the case of CSW, developers are able to select the metadata repository more suitable to their preferences/requirements. However, the OGC CSW standard defines operations, requests and metadata formats that should be supported. For instance, queries submitted to a CSW should be formatted as Filter Encoding or as CQL. The former is a XML encoded query language, while the later is a human readable text encoded query language (OSGeo, 2012)(Vretanos, 2005).

Most common implementations of CSW use a relational database as the metadata records repository. For instance, GeoNetwork currently one of the most popular CSW implementations, uses by default a McKoiDB relational database, although it can connect to MySQL, PostGreSQL and other RDBMS (Dunne et al., 2012). Because of the nature of the metadata repository, currently queries are performed

by matching strings to selected metadata elements.

In this paper we propose the use of semantic web technologies to store and query metadata records. By using these technologies we are able to take advantage of inference and reasoning mechanisms not available on relational databases. In Section 2 we review research conducted by other teams in the same field. In Section 3 we describe how we have implemented our model. Finally, in Section 4 we present our conclusions and outline future research.

2 RELATED RESEARCH

Spatial information is offered by different providers through web services that implement standards such as WCS, SOS, WFS or WMS. The deployed services might have heterogeneous characteristics regarding software components, languages, or providers. However, by implementing OGC standards all of them have common request and response contents, parameters and encodings. These common elements allow a user to access different services using a proven, safe strategy . In order to allow datasets to be *discoverable*, they have to be published in a catalogue service that implements the CSW standard. The metadata for the datasets is obtained by the catalogue service with a *harvest* operation. The user in order to discover a specific dataset, submits a query. The server processes the query using a *string matching process*, and sends a response to the user. Once the user has identified the relevant dataset, she is able to obtain the data and perform a specific analysis.

The *string matching process* is a major limitation in the current SDI. This limitation has been previously identified by other researchers. In (Kammersell and Dean, 2007) the authors aim to integrate heterogeneous datasources. In this research the authors propose the creation of a layer that translates the users query formulated in OWL into WFS XML request format. Later, they propose do the inverse process with the results. Another approach is proposed by (Kolas et al., 2005). Here the authors propose the implementation of five different ontologies: 1) *Base Geospatial Ontology* for basic geospatial concepts resulting from the conversion of GML schemas into OWL. 2) *Domain Ontology*, this is the user‘s ontology. Its purpose is to link user‘s concepts to the Base Geospatial Ontology. 3) *Geospatial Service Ontology*, used to describe services and allow discovery. 4) *Geospatial Filter Ontology*, which is used to formalize filter description and use. 5) *Feature Data Source Ontology*, to represent the characteristics of the features returned from the WFS. Another approach is described by (Harbelot et al., 2013), here the authors suggest the integration of data from OGC services into a triplestore with a focus on the WFS filters. In (Janowicz et al., 2010) (Janowicz et al., 2012), the authors propose the addition of semantic annotations for each level of a geospatial semantic chain process that involves OGC services. For instance, they propose specific semantic annotations at the level of the service OGC Capabilities document that would correspond to all the datasets managed by the service. Other annotations would correspond to specific data layers. Spatial Data with semantic annotations could later be processed and semantically analysed using custom made reasoning services. To achieve this goal they propose the deployment of OGC services capable of interacting with libraries such as *Sapience* which would result in richer data and data descriptions. However there is little development in this direction. At the moment there is little use of semantic annotations on OGC capabilities documents.

In (Gwenzi, 2010) the author describes the CSW limitations by evaluating GeoNetwork. According to the author there are three possible ways to add semantic capabilities to the CSW: 1) Linking keywords to concepts in the *getCapabilities* response. 2) By adding an ontology browser to the GeoNetwork client interface. 3) Using ebRIM extensions to add ontologies to the CSW. In (Gwenzi, 2010) the author implements the third option.

Another experience is presented by Yue et al. (2006). In this work the authors extend the ebRIM CSW specification by: 1) Adding new classes based to existing ebRIM classes; and 2) Adding Slots to existing classes, thus creating new attributes. With these additions the authors are able to store richer metadata records in the catalogue. The authors identified two options to implement an upgraded search functionality: 1) Create an external component without further modification of the CSW schemas; 2) Modify the CSW adding semantic functionalities to the existing CSW schemas. In (Yue et al., 2006) the authors opted for the first option. Yue et al. (2011) extends this work, focusing on geoservices (Yue et al., 2011).

A different approach is used by (Lopez-Pellicer et al., 2010). In this research the goal is to provide access to data stored in CSW as Linked Data. In order to achieve this goal the authors developed CSW2LD, a middle layer on top of a conventional CSW based server. It allows the server to mimic other Linked Data sources and publish metadata records. CSW2LD wraps the following CSW requests: *GetCapabilities*, *GetRecords* and *GetRecordById*.

A very interesting work in progress is described in (Pigot, 2012). This is a website describing a proposal

by a team from the GeoNetwork developer community. The authors intend to perform a major change in GeoNetwork, allowing it to store metadata as RDF facts stored in a RDF repository. They intent to use SPARQL/GeoSPARQL to retrieve data. The website describe technical characteristics of GeoNetwork and mentions fields that require work in order to implement the project. Currently queries in GeoNetwork are formatted as Filter Encoding or as CQL. Any implementation of a RDF metadata repository would need to consider a translation mechanism between the current queries format to SPARQL (a W3C recommendation) (DuCharme, 2011). Currently GeoNetwork handles spatial constraints using GeoTools. In the semantic web domain, spatial queries are performed using GeoSPARQL (Kolas and Batle, 2012). According to the authors it is not clear if GeoSPARQL is mature enough to handle metadata spatial queries. Even more there is no mechanism to translate spatial constraints into GeoSPARQL. Regardless of the advantages that semantic web technologies might bring into CSWs there is scarce research on this topic. By the time we wrote this paper, there was no further development in (Pigot, 2012) and the website was last updated by the end of October of 2012.

3 IMPLEMENTATION

A regular implementation of OGC CSW, works as a web service that communicates with a data repository that stores metadata records. According to the OGC standard, the catalog should accept requests formatted as *Filter Encoding*, which is a XML based language, designed to express queries. The web service, translates these queries into a suitable format, such that it can communicate with its data repository. In most of the implementations of CSW, the data repository is relational database.

In this paper, we present a proof of concept implementation, designed to show the benefits of using ontologies in a geospatial data catalog service. We have developed a minimalistic implementation of the CSW standard. A major difference between our system and traditional implementations is that we use a triplestore as our metadata repository. We opted for a Parliament triplestore, because of its spatial capabilities thanks to its support for GeoSPARQL. We developed an ontology in the triplestore, and mapped the metadata records to instances of classes specified in the ontology. Thanks to this, it is possible to use *superclass - subclass* relationships in the metadata search process.

In traditional CSW implementations, a spatial search uses the values of the bounding box of the dataset, as specified in the metadata record. However, the number of users with an understanding of coordinate systems, good enough to allow them to search datasets using only the bounding box coordinate values is quite limited. In our approach, we implement an ontology class called *ToponymUnit*. Instances of this class are geographic features with labels familiar to the users. In our implementation, the user can search for metadata records whose bounding box has specific spatial relations with instances of the class *ToponymUnit*. Using this approach, users can submit queries such as: retrieve metadata records that are *within* the toponym unit known as *France*.

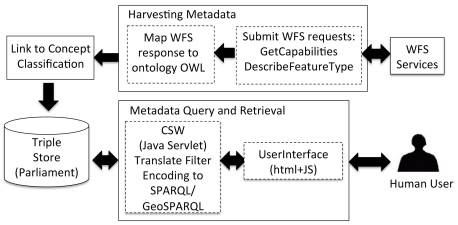


Figure 1: Architecture implementation.

Our current implementation is able to respond to standard *GetRecords* requests submitted as *POST*. The response of our system follows the *csw:SummaryRecord* format. A crucial part of the implementation is the translation of requests from *Filter Encoding* to *SPARQL/GeoSPARQL*.

Figure 1 depicts the processes in the proposed system. In the next subsections we further describe how we obtain the information necessary to construct the metadata records, how we map this information to an ontology, and finally how we perform queries.

3.1 Harvesting Metadata

We focus our research on metadata records for datasets available through services that implement the OGC Web Feature Service (WFS) standard. A service that implements WFS can contain one or many datasets. We identify the datasets available on a WFS and create a metadata record for each one of them. The metadata record is stored as an instance of the class *abc:MetadataRecord* .

In our ontology we implement a class called *geo:Feature* that represent features with spatial representation. Although, a metadata record is an abstract description, it does have a spatial component represented by the bounding box of the dataset it describes. Due to the spatial nature of the metadata records, we define the class *abc:MetadataRecord* as a subclass of *geo:Feature* (See Figure 2).

In our ontology we implemented Dublin Core elements, as properties for instances of the class

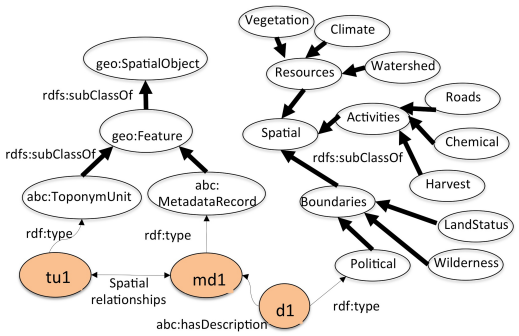


Figure 2: Classes, instances and relationships in the proposed model.

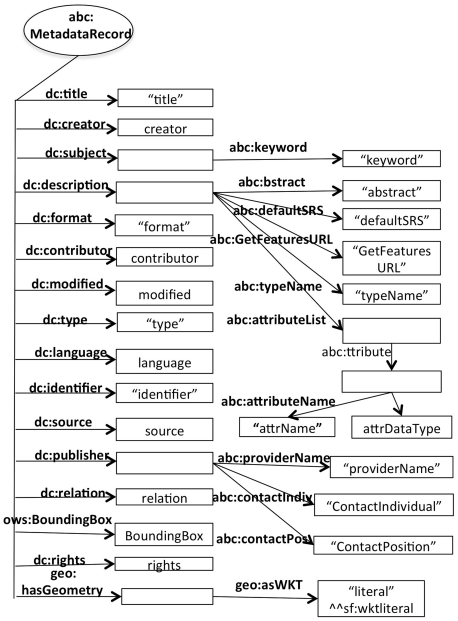


Figure 3: Properties of instances of the class *abc:MetadataRecord*

abc:MetadataRecord. We have developed a harvest tool that queries the WFS and constructs triples with the responses. Our tool is a Java application that makes use of two requests that are part of the core OGC WFS standard: *GetCapabilities* and *DescribeFeatureType*. From the *GetCapabilities* request, we obtain a general description of the catalog service and information regarding the available datasets on it. With the *DescribeFeatureType* request, we can obtain the list of attributes for the actual features that compose the dataset. We have tested our tool with 17 services that implement WFS, having as a result 2690 metadata records. Figure 3 depicts the Dublin Core elements as properties of the class *abc:MetadataRecord*.

The mapping process between the WFS request responses and ontology properties is done with our harvest tool. The WFS respond to requests with a XML document. Our harvest tool, navigates through the response document, identifies relevant elements and maps them to specific properties in our ontology (See Figures 1 and 3).

For instance, the following code is part of a *GetCapabilities* response containing information regarding the WFS publishing entity.

```
 <ows:ServiceProvider>
<ows:ProviderName>
Provider X
</ows:ProviderName>
<ows:ServiceContact>
<ows:IndividualName>Helbert<ows:IndividualName/>
<ows:PositionName>GIS Manager<ows:PositionName/>
</ows:ServiceContact>
</ows:ServiceProvider>
```

This information is mapped into the *dc:publisher* Dublin Core element (See Figure 3).

```
abc:md1 a abc:MetadataRecord .
abc:md1 dc:publisher _:B01 .
_:B01 abc:ProviderName "Provider X" .
_:B01 abc:PositionName "GIS Manager" .
_:B01 abc:IndividualName "Helbert" .
```

The following XML code, is another part of a *GetCapabilities* response.

```
 <FeatureType xmlns:example=
 "http://www.example-provider.org/example">
<Name>example:name</Name>
<Title>Example dataset title</Title>
<Abstract>Example abstract</Abstract>
<ows:Keywords>
<ows:Keyword>example keyword1</ows:Keyword>
</ows:Keywords>
<DefaultSRS>
urn:x-ogc:def:crs:EPSG:4326
</DefaultSRS>
<ows:WGS84BoundingBox>
<ows:LowerCorner>-5.84 37.75</ows:LowerCorner>
<ows:UpperCorner>11.02 54.63</ows:UpperCorner>
</ows:WGS84BoundingBox>
</FeatureType>
```

From this segment of the response, we can obtain information for: *dc:title*, *dc:subject*, *dc:description* and *ows:BoundingBox*.

```
abc:md1 a abc:MetadataRecord .
abc:md1 dc:title "Example dataset title" .
abc:md1 dc:subject _:B02 .
_:B02 abc:keyword "example keyword1" .
abc:md1 dc:description _:B03 .
_:B03 abc:abstract "Example abstract" .
_:B03 abc:defaultSRS "EPSG:4326" .
abc:md1 geo:hasGeometry _:B04 .
_:B04 geo:asWKT "POLYGON((-5.84 37.75,
-5.84 54.63, 11.02 54.63, 11.02 37.75,
-5.84 37.75))"^^sf:wktLiteral .
```

Additionally, our harvesting tool, submits a *DescribeFeatureType* request for each layer of information found in the WFS. From the response we are able to obtain a list of attributes for the spatial features that compose the dataset. The following XML code depicts part of a response to a *DescribeFeatureType* request.

```
<xsd:complexType name="country_boundsType">
<xsd:complexContent>
<xsd:extension base="gml:AbstractFeatureType">
<xsd:sequence>
<xsd:element maxOccurs="1" minOccurs="0"
name="THE_GEOM" nillable="true"
type="gml:MultiSurfacePropertyType"/>
<xsd:element maxOccurs="1" minOccurs="0"
name="AREA" nillable="true"
type="xsd:double"/>
</xsd:sequence>
</xsd:extension>
</xsd:complexContent>
</xsd:complexType>
```

Our harvesting tool, browses through the response, identifies relevant elements and create an attribute list within the dublin core element *dc:description* , and creates the following triples:

```
abc:md1 a abc:MetadataRecord .
abc:md1 dc:description _:B03 .
_:B03 abc:attributeList _:B05 .
_:B05 abc:attribute _:B06 .
_:B06 abc:attributeName "THE_GEOM" .
_:B06 abc:attributeDataType
gml:SurfacePropertyType .
_:B05 abc:attribute B07 .
_:B07 abc:attributeName "AREA" .
_:B07 abc:attributeDataType xsd:double .
```

3.2 Link to Concept Classification

By using semantic web technologies we are not limited to string matching queries. We can also use inference mechanisms based on subsuming and established relationships between terminology and concepts. To test these capabilities, we have implemented a taxonomy with *domain ontology* classes. The relationships between these concepts are of the type *subclassOf*. Individual datasets are represented as instances of the domain ontology classes. Each dataset is described by a metadata record, which is an instance of the class *abc:MetadataRecord* . The link between these two instances is given by the property *abc:hasDescription* (See Figure 2). The following triples depict the relation between an instance of *abc:MetadataRecord* and an instance of the domain ontology class *abc:Political*.

```
abc:d1 a abc:Political .
abc:md1 a abc:MetadataRecord .
abc:d1 abc:hasDescription abc:md1 .
```

Figure 2 depicts our proposed domain ontology. Using the *domain ontology* we can make inferences regarding the class membership. From the previous example, because $abc : Political$ is a subclass of $abc : Boundaries$ and $abc : Spatial$, we can infer that $abc : md1$ is also a member of these two classes.

The goal of this paper is to show potential uses of semantic web technologies for metadata record management. At the moment our focus is not on the automatic classification of datasets within a domain ontology. This task can be achieved with a variety of methods for instance: Naive Bayes, Decision Rules, Neural Networks, among others. For this experiment, we decided to create an instance of the class $abc : Spatial$ for each metadata record, and later add further specification of the instance randomly, only in order to test the system query capabilities. Our future development plans include the implementation of sophisticated methods for the dataset classification. An interesting work in this field, although not in the spatial domain is (Werner et al., 2012).

3.3 Toponym Elements

In order to facilitate the identification of suitable spatial datasets we define the class *abc:ToponymUnit* as a subclass of *geo:Feature* . Instances of the class *abc:ToponymUnit* are geographic features with known, accepted names. Our system enables users to integrate into their metadata queries spatial relations between the spatial components of metadata records (bounding box) and toponym instances (See Figure 2).

To populate the class *abc:ToponymUnit*, we use a country political boundaries dataset from Esri and DeLorme Publishing Company, Inc. under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License (ESRI, 2011). The dataset is a shapefile with 668 multipolygon features. Our goal, is to create instances of the class *abc:ToponymUnit*, each instance representing an area with a known political designation.

However, before translating the political boundaries dataset into triples, it was necessary to perform the following steps: 1) Convert the multipolygon features to polygon ones. 2) Delete polygons that we considered too small for practical purposes. 3) Simplify the remaining polygons by reducing the number of vertices. 4) Translate the political boundaries dataset into instances and triples using a customized Java program, implemented with Jena and GeoTools libraries. and finally 5) Upload the triples into our triplestore. The pre-loading processing was done using Quantum GIS and GeoTools. The final result is

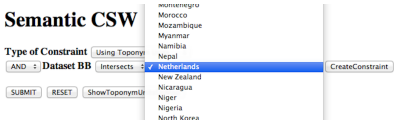


Figure 4: HTML user interface: Defining a constraint using a spatial relationship with an instance of th class *abc:ToponymUnit*

3037 instances of the class *abc:ToponymUnit*.

3.4 Metadata Query

In order to test our implementation, we need to enable users to submit queries to the catalog service and browse through the query results. This process is composed by the following four sub-processes : 1) Definition of the query. 2) Mapping the user request into a suitable metadata repository query format. 3) Mapping the triplestore response into an allowed OGC standard. 4) The visualization of the results in the user interface (See Figure 1).

3.4.1 Definition of the Query

The OGC standard query language for services implementing CSW is called Filter Encoding. This is a language based on XML . Queries with Filter Encoding might become long XML documents that require a strict syntax, therefore it is necessary a software application that helps the user to compose them. In our system this task is done in a user interface deployed as a web page. It uses a combination of HTML and JavaScript, to enable users to compose complex queries.

In order to help the user to compose queries, the web site requests from the Server: 1) The list of domain ontology classes, 2) The list of labels associated to instances of *abc* : *ToponymUnit*. It uses this information to populate combo boxes, allowing the user to compose constraints (See Figure 4).

The JavaScript application receives the user input and formats it as a XML query document following the Filter Encoding specification. The application allows multiple constraints to be linked using the operators *AND* and *OR*. When the query is completed the user submits it as a *POST*. The following XML code represents a query as formatted by the JavaScript running on the website.

```
<GetRecords><Query>
<Constraint><Filter><And>
<PropertyIsLike>
<ValueReference>dc:title</ValueReference>
<Literal>water</Literal>
</PropertyIsLike>
<DescribesInstanceOf>abc:Harvest
</DescribesInstanceOf>
<sfIntersects>
<ValueReference>BoundingBox</ValueReference>
<ToponymUnit>Netherlands</ToponymUnit>
</sfIntersects>
</And></Filter></Constraint>
</Query></GetRecords>
```

3.4.2 From Filter Encoding to SPARQL/GeoSPARQL

We are using as our metadata repository a triplestore, therefore it is necessary to translate request from Filter Encoding to SPARQL/GeoSPARQL. This task is accomplished by our Servlet application. Once the XML query arrives, the application proceeds to decompose it, into its constituent constraints.

In a SPARQL query, we distinguish three components. 1) The specific elements or nodes we are requesting; 2) A set of triples that define a pattern the SPARQL engine is going to look for; and 3) The filter component, where we define a set of boolean value conditions for the triples that match the previously defined pattern.

Our servlet application translates each constraint separately into the respective set of triples pattern and filter conditions. Using the interface the user is able to define three types of constraints:

1. Alphanumeric attributes in the metadata record: The user can select one attribute in the metadata record, and perform a string matching, using the operators *PropertyIsEqualTo* and *PropertyIsLike*. In the later case the SPARQL implementation will require the definition of a filter component:

   ```
   <PropertyIsLike>
   <ValueReference>dc:title</ValueReference>
   <Literal>water</Literal>
   </PropertyIsLike>
   ```

   would be translated to the triple pattern:

   ```
   ?md a abc:MetadataRecord.
   ?md dc:title ?xTitle.
   ```

   with the filter component:

   ```
   (regex(?xTitle,"water","i"))
   ```

2. Domain ontology class membership: Each metadata record describes an entity with a class membership. This type of constraint allows the user to identify the class membership of the entity. For example, the constraint:

   ```
   <DescribesInstanceOf>abc:Harvest
   </DescribesInstanceOf>
   ```

   Is translated as the SPARQL triples:

   ```
   ?md a abc:MetadataRecord.
   ?ds abc:hasDescription ?md.
   ?ds a abc:Harvest.
   ```

3. Using toponym elements and spatial relationships: In this case the user identifies a toponym of interest, then she defines a spatial relationship between the bounding box of the metadata record and the geometry of the selected toponym. We implement the spatial operators *Disjoint*,*Intersects*, *Contains* and *Within* (See Figure 4). For example, the following XML extract, indicates that the metadata records bounding box should intersect the geometry of *Netherlands*:

```
<sfIntersects>
<ValueReference>BoundingBox</ValueReference>
<ToponymUnit>Netherlands</ToponymUnit>
</sfIntersects>
```

The constraint is translated as the following triple pattern:

```
?md a abc:MetadataRecord.
?md geo:hasGeometry ?boundingbox.
?boundingbox geo:asWKT ?boundingbox_wkt.
?topoUnit a abc:ToponymUnit.
?topoUnit abc:CountryName "Netherlands".
?topoUnit geo:hasGeometry ?topoGeo.
?topoGeo geo:asWKT ?topoWKT.
```

plus the additional filter component:

```
(geof:sfIntersects
(?boundingbox_wkt,?topoWKT))
```

Once all the constraints have been translated, the triples and filter elements are merged into a syntactically correct SPARQL/GeoSPARQL query, which is submitted to the triplestore. The following code, depicts the SPARQL query resulting from combining all the triple patterns and filter components from the previous examples:

```
SELECT DISTINCT ?md ?xTitle
Where
{?md a abc:MetadataRecord.
?md dc:title ?xTitle.
?md geo:hasGeometry ?boundingbox.
?boundingbox geo:asWKT ?boundingbox_wkt.
?ds abc:hasDescription ?md.
?ds a abc:Harvest.
?topoUnit a abc:ToponymUnit.
?topoUnit abc:CountryName "Netherlands".
?topoUnit geo:hasGeometry ?topoGeo.
?topoGeo geo:asWKT ?topoWKT.
FILTER((regex(?xTitle,"water","i"))&&
(geof:sfIntersects(?boundingbox_wkt,?topoWKT)))}
```

The response from the triplestore is then formatted by the servlet as *csw:SummaryRecord* and sent to the client website. The results are visualized in the website allowing the user to examine the metadata records (See Figure 5).

3.5 Smart Queries

A *smart query* requires the combination of diverse datasources as described in (Goodwin, 2005). However, first the researcher must be able to identify the most suitable dataset for the analysis. Our implementation aims to help users in this task. By using a domain ontology, we improve the user's query capabilities. Our use of toponyms, allows the user to select areas of interest by name, and establish specific spatial relationships with the dataset of interest. The actual features of the dataset can later be obtained using the value of *abc:GetFeaturesURL*, in the metadata record.

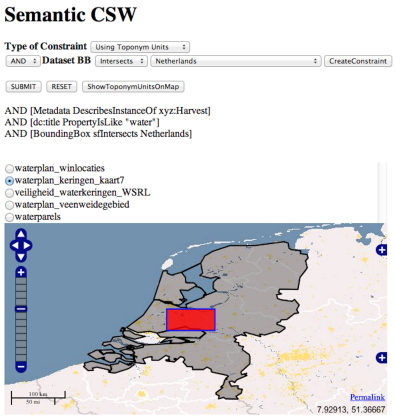


Figure 5: HTML user interface showing the results of the example query.

4 CONCLUSIONS

In this work we present a simplified CSW implementation with a triplestore as a metadata repository. Our implementation has a working translator that is able to convert Filter Encoding queries into SPARQL/GeoSPARQL ones. The system allows complex queries that can take advantage of inference mechanisms provided by Semantic Web technologies.

At this point, our system only uses inference based on class to subclass relationships. However, we plan to extend these capabilities to include relationships between concepts and automatic class membership determination using a domain ontology.

Our approach to capture metadata information is generic, takes advantage of the OGC standard interfaces. With our harvesting tool we were able to create 2690 metadata records. However the information supplied by the WFS publishing entities has limitations and is in many cases incomplete. Our metadata records contain 1384 distinct keywords including 383 actual URLs. However in no case the URLs referred to any ontology or formalized vocabulary. From the URLs, 217 were links to html documents, and 52 to XML documents. In both cases the documents contained extended metadata descriptions of the datasets. All

the datasets with URL of extended descriptions were provided by one single WFS deployment (giswebservices.massgis.state.ma.us/geoserver/wfs?), the rest of the keywords were strings with no formal semantics associated. Our metadata harvesting tool also obtained information regarding the names of the attributes of the dataset. In total we have obtained 6331 individual attribute names, all of them were strings with no formal semantics associated.

The use of extended descriptions in XML and HTML documents is not a standard practice among the WFS publishing entities. However, in case we find more documents of this kind, we can upgrade the harvesting tool in order to allow it to get information from the associated documents.

The results of our current implementation are promising, in the near future we will implement an automatic classification of metadata records based on harvested metadata using a domain ontology.

ACKNOWLEDGEMENTS

This research is supported by: 1) Conseil régional de Bourgogne. 2) Direction Générale de l'Armement, see: http://www.defense.gouv.fr/dga/.

REFERENCES

DuCharme, B. (2011). *Learning SPARQL*. O'Reilly Media, Inc.

Dunne, D., Leadbetter, A., and Lassoued, Y. (2012). ICAN Semantic Interoperability Cookbooks. Technical report, International Coastal Atlas Network.

ESRI (2010). GIS Best Practices: Spatial Data Infrastructure (SDI). http://www.esri.com/library/bestpractices/spatial-data-infrastructure.pdf. Accessed: July 2013.

ESRI, D. (2011). World administrative units. http://resources.arcgis.com/content/data-maps/10.0/world. Accessed on May 2013.

Goodwin, J. (2005). What have ontologies ever done for us - potential applications at a national mapping agency. In *OWL: Experiences and Directions (OWLED)*.

Gwenzi, J. (2010). Enhancing spatial web seach with semantic web technology and metadata visualization. Master of science, University of Twente.

Harbelot, B., Arenas, H., and Cruz, C. (2013). Semantics for Spatio-Temporal "Smart Queries". In *Poster presentation in the 9th. International Conference on Web Information Systems and Technologies*, Aachen, Germany.

Janowicz, K., Schade, S., Broring, A., Kebler, C., Maue, P., and Stasch, C. (2010). Semantic Enablement for Spatial Data Infrastructures. *Transactions in GIS*, 14(2):111–129.

Janowicz, K., Scheider, S., Pehel, T., and Hart, G. (2012). Geospatial Semantics and Linked Spatiotemporal Data - Past, Present and Future. *Semantic Web - Interoperability, Usability and Applicability*, 3(4):1–10.

Kammersell, W. and Dean, M. (2007). Conceptual Search: Incorporating Geospatial Data into Semantic Queries. In Scharl, A. and Tochtermann, K., editors, *The Geospatial Web*, Advanced Information and Knowledge Processing, pages 47–54. Springer London. 10.1007/978-1-84628-827-2_5.

Kolas, D. and Batle, R. (2012). GeoSPARQL User Guide. http://ontolog.cim3.net/file/work/SOCoP/Educational/GeoSPARQL User Guide.docx Accessed on May 2013.

Kolas, D., Hebeler, J., and Dean, M. (2005). Geospatial Semantic Web: Architecture of Ontologies. pages 183–194.

Lopez-Pellicer, F. J., Florczyk, A., Renteria-Aguaviva, W., Nogueras-Iso, J., and Muro-Medrano, P. R. (2010). CSW2LD: a Linked Data frontend for CSW.

OGC (2012). OGC Institutional Web Site. http://www.opengeospatial.org/. Accessed: September 2013.

OSGeo (2012). CQL. http://docs.geotools.org/latest/userguide/library/cql/cql.html. Accessed on November 2012.

Pigot, S. (2012). Using RDF as Metadata Storage. http://trac.osgeo.org/geonetwork/wiki/rdfstore. Accessed on May 2013.

Vretanos, P. A. (2005). Filter Encoding Implementation Specification. online. Accessed on May 2013.

Werner, D., Cruz, C., and Nicolle, C. (2012). Ontology-based Recommender System of Economic Articles. In *WEBIST 2012*, pages 725–728.

Yue, P., Di, L., Yang, W., Yu, G., and Zhao, P. (2006). Path planning for chaining geospatial web services. In *Proceedings of the 6th international conference on Web and Wireless Geographical Information Systems*, W2GIS'06, pages 214–226, Hong Kong, China. Springer-Verlag.

Yue, P., Gong, J., Di, L., He, L., and Wei, Y. (2011). Integrating Semantic Web Technologies and Geospatial Catalog Services for Geospatial Information Discovery and Processing in Cyberinfrastructure. *GeoInformatica*, 15:273–303. 10.1007/s10707-009-0096-1.

Testing Discovered Web Services Automatically

Pinar Karagoz[1] and Selma Utku[2]
[1]*Department of Computer Eng, Middle East Technical University, Ankara, Turkey*
[2]*Aselsan Inc, Ankara, Turkey*
karagoz@cengmetu.edu.tr, sutku@aselsan.com.tr

Keywords: Web Service, Web Service Testing, Semantic Dependency Analysis, Mutation Analysis.

Abstract: The reliability of web services is important for both users and software developers. In order to guarantee the reliability of the web services that are invoked and integrated at runtime, mechanisms for automatic testing of web services are needed. A basic issue in web service testing is to be able to generate appropriate input values for web services and to estimate whether the output obtained is proper for the functionality. In this work, we propose a method for automatic web service testing that uses semantics dependency-based and data mutation-based techniques to generate different test cases and to analyze web services. We check whether the services function properly under the input values generated and enriched from various data sources and we check robustness of web services by generating random and erronous data inputs. Experimental evaluation with real web services show that proposed mechanisms provide promising results for automatic testing of web services.

1 INTRODUCTION

Web services provide interaction between different distributed applications and the use of web services becomes one of the most preferable technologies by software developers and web users. They are published in high numbers, they become outdated very rapidly, and there is no standard control mechanism for their reliability. Traditional offline and manual testing processes are not always applicable to testing of web services. Therefore, it is necessary to have a mechanism for online and automated testing of dynamically discovered and selected web services to be included in a software application (Wang et al., 2007), (Dranidis et al., 2007).

In this work, we present a method for automated testing of web services. An important issue that we target in this work is to generate appropriate input values for web services automatically. To this aim, we adopt two previous techniques from different areas: mutation analysis, and semantic analysis. In mutation analysis, a web service is tested by using random and specified values in different ranges that are set according to the parameter type. Whereas in semantic analysis, dependencies among web services that are published by the same service provider are semantically analysed. Different test cases are generated and the test score of a web service is obtained by examining the outputs in two dimensions. Since we do not know the exact behaviour of the services, we aim to employ simple yet effective methods here. In the first dimension, if an exception occurs during the execution of the web service request, the web service is considered unsuccessful. The second dimension is based on checking whether the web service returns different results for different test cases.

In this study, we work on web services specified in Web Services Description Language (WSDL)[1], which is the prominent specification method for web services. The methods proposed in this work do not need semantic description of services. The reason behind this choice is based on the observation that almost all of the published web services do not have semantic definitions. One basic assumption is that web services have providers and they publish a set of services, and these service are generally related such that they belong to the same domain and output of one service may be input for another one. We use such dependencies in semantic analysis of web services. The proposed method is implemented in a domain specific web service discovery system. The web service testing module in this system is in charge of checking whether the discovered services function correctly. The performance of the proposed method is evaluated on synthetic and real world web services.

[1]http://www.w3.org/TR/wsdl

The experimental results indicate the usability of the approach.

The main contribution of this work in comparison to previous work can be summarized as follows: The proposed method does not rely on an external mechanism or framework, it can be easily applied. It does not assume availability of semantic service description or description of internal mechanism of web service. For input generation for web services, it combines data mutation based technique with semantic dependency analysis among web services. Test score generation process is performed automatically.

The rest of this paper is organized as follows. In Section 2, related work is summarized. In Section 3, general architecture of the proposed web service testing method is described. In Section 4, mutation-based analysis is presented. In Section 5, semantic based analysis is given and this is followed by the score generation described in Section 6. Evaluation results are presented in Section 7. The paper is concluded in Section 8 with an overview and future work.

2 RELATED WORK

Before using the web service by service consumers, testing is needed to guarantee the correctness and robustness of web services. The study of Martin et al. (Martin et al., 2006) is one of studies that emphasize the robustness testing of web services by using WSDL. It presents a framework to generate and execute robustness test cases automatically.

In (Bai et al., 2005), Bai et al. propose an approach about WSDL-based test generation and test case documentation to provide reusability of generated test cases. Test cases are generated in four levels: test data generation, individual test operation generation, operation flow generation, and test specification generation. Test data is generated by analyzing WSDL message definitions. In individual test operation generation level, input parameters of web services are analyzed and test operations are generated. In operation flow generation level, the sequence of the web services are determined by the analysis of dependencies between the web services. In this approach, three dependencies are used; input dependency, input/output dependency and output dependency.

Siblini et al. (Siblini and Mansour, 2005) propose a mutation testing method for web services. The aim of this approach is to find errors relevant to both the WSDL interface and the logic of web service programming. In their work, mutant operators to the WSDL document of web services are defined and mutated web service interfaces are generated. With each modification, a new version of test case is created and it is called a mutant. Mutant operators are applied to input parameters and output parameter of web services and the data types that are defined in the WSDL document. .

Obtaining valid inputs and outputs is a tedious work and often such information is not readily available. AbuJarour et al. (AbuJarour and Oergel, 2011) propose an approach to generate annotations for web services, i.e.., valid input parameters, examples of expected outputs and tags, by sampling invocations of web services automatically. The generated annotations are integrated to web forms to help service consumers for actual service invocations. In this approach, in order to generate valid parameters, various resources such as random values, outputs of other web services that are provided by the same web service provider and different providers, external data sources such as WordNet, DBpedia, Freebase, are used.

Although the proposed method have similarities with (Bai et al., 2005), (Siblini and Mansour, 2005) and (AbuJarour and Oergel, 2011), as the basic difference, in this study, the emphasis is on generating appropriate inputs for testing and generate an overall test score automatically.

3 GENERAL ARCHITECTURE

The proposed work aims to automatically test web services that are specified in WSDL. Each service provider has a WSDL document to specify the information about the provided web services. This document contains the names of the web services, the attributes of input and output parameters of the web services and also user defined types. This document provides valuable information for service invocation. However, this information is not sufficient to test the web services. Since the behavioral information of web service is not available, only black-box testing can be performed for validation and testing of web service.

Within the scope of this work, web service validation is defined as the process of checking whether the web service is still alive and accessible or not by invoking the web service by simple appropriate parameter values. On the other hand, web service testing is the task of checking whether a web service functions as it should be. Both processes require generation of input values for test cases. To this aim, firstly, the types of the input and output parameters of a given web service should be identified. As the atomic parameter types, boolean, character, integer and floating point number are the most common ones. By using

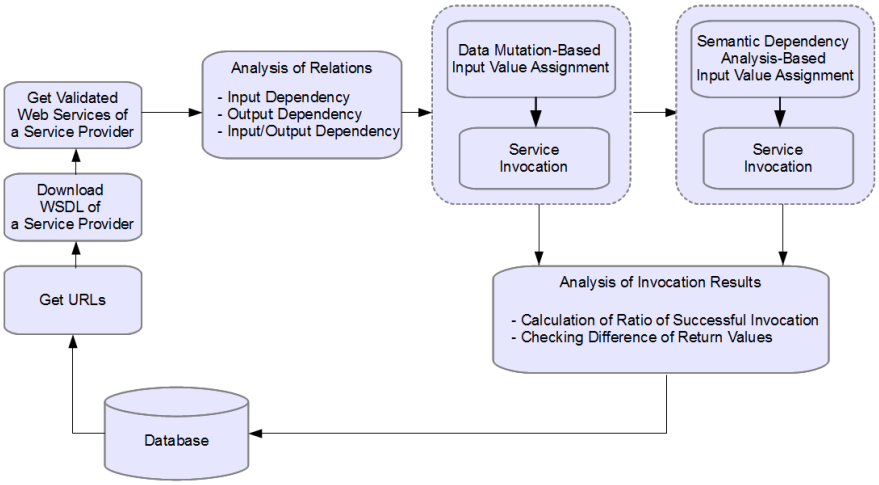


Figure 1: General Architecture of Web Testing Approach.

the atomic types, user-defined complex types can be constructed. To analyze and generate a value for basic types is more straightforward than that of complex types. In order to generate test data of complex data types, the structure is recursively analyzed until it encounters with basic data types. Since the aim of the validation process is to check whether the web service is still alive and accessible or not by invoking the web service, setting simple default values to input parameters is sufficient for validation process.

In order to test a web service, one criteria is that no exception is taken when the web service is invoked by using the generated input values. Another important point is to check whether the service generates different outputs for different input values. If a service always returns the same value in spite of different input parameters, it is considered that it is a test service or this service has no implementation. Such services fail the test.

The general architecture of the proposed approach is shown in Figure 1. Web service testing process starts with getting the URLs of service providers. At this point, we assume that the service is already discovered, and hence service provider's address is available. By downloading and analyzing the content of WSDL documents, information as to which web services are provided and how to invoke are obtained. For each provided web service, the validation process is performed in order to check whether it is alive or not. On the validated services set, input dependency, output dependency and input/output dependency relations are analyzed.

The next step is test case generation with data mutation-based method. By this method, the values of input parameters are generated according to the parameter type declarations in the service description. With the generated input parameters, the web service is invoked and the invocation result is saved for score calculation. Following this, test case generation with semantic dependency analysis is performed. For each web service, the values for input parameters of web service are determined according to the semantic dependency analysis and the web service is invoked with the generated input parameters. As in the previous step, for each invocation, the results are saved. Finally, for each web service, the final test result is calculated on the basis of the results of all invocations.

4 MUTATION ANALYSIS

In software engineering, the purpose of the mutation testing is to help the tester develop effective tests or locate weaknesses in the test data used for the program or in sections of the code that are seldom or never accessed during execution. The proposed method is inspired from mutation testing methods, however the aim is quite different. The aim of mutation testing is to measure test adequacy. On the other hand, the aim of data mutation in the proposed method is to generate test cases. In traditional mutation testing, mutation operators are used to transform the program under test. In contrast, this method is applied for generating random and erroneous data inputs.

In mutation-based test case generation, a parameter can have different values in the range of its type domain. In this method, the various random and error-prone data input values are generated and they are grouped. When we use the generated values in each

range for input parameter, we expect that invoked web service presents different behaviours.

We mention that types of input parameters can be fundamental or complex type. In case of complex type, the generator recursively analyzes the structure of the input parameter type until it reaches the fundamental type. Various values in the range of its type are generated and the generated values are grouped. For each group, the range of parameter value is also constrained. We apply several versions of data nutation. In Mutant Version 0, positive values are generated for input parameters. When the web service is invoked by using these values, it is expected that invoked web service returns a proper value. In Mutant Version 1, 2, 3, 4, 5 and 6, erroneous input values are generated.

For numeric types, in Mutant Version 1, input parameter is set to 0. By this way, it can be checked whether the web service prevents division by zero error or pointer address error. In Mutant Version 2, the numeric input parameter is set to -1 and thus it is checked whether unsigned to signed conversion error, out-of-bounds memory access error, signed/unsigned mismatch warning in comparison is prevented by web service. Incorrect sign conversions can lead to undefined behaviour and the web service can be crashed. In Mutant Version 3 and 4, boundary values are generated. In Mutant Version 5 and 6, very high and low values are generated randomly. Thus, the fault resistance of the web service is checked. For each type, the values that can be generated are different. Therefore, the mutant groups are constructed differently with respect to the parameter type.

For String typed parameters, it is very hard to generate valid values. Although the web services are not annotated, if we know the domain of the web services, it is possible that the instances can be directly taken from the ontology or the ontology can be populated with instances by using public resources. As the first step, the name of input parameters are semantically compared with the ontology terms in order to find the similarity value and ontological position of the input parameters. We obtain the ontology terms with matching degree above the threshold value. We check whether there is any instance for the ontology term corresponding to the name of input parameter. If it has at least one instance, a random one of them is used for input parameter value in test case generation. Otherwise, the hyponym terms of the name of input parameter are searched and if any term is obtained, the instances of the obtained hyponym terms are used, if available.

5 SEMANTIC DEPENDENCY ANALYSIS

A service provider generally publish multiple services and some of them have interactions with each other. In such a situation, for testing atomic web services, test cases can be generated by using these interactions. Semantic dependency analysis considers the following three types of dependencies:

- Input dependency: A web service WS1 is input dependent on WS2 if and only if WS1 and WS2 share at least one input parameter that has the same type and same name.
- Output dependency: A web service WS1 is output dependent on WS2 if and only if output parameters of WS1 and WS2 have the same type and the same name.
- Input/output dependency: A web service WS1 is input/output dependent on WS2 if and only if at least one input parameter of WS1 has the same type and similar name with at least one field of output parameter of WS2.

In this work, especially, *semantic input/output dependency* is used. The generation of the values for input parameters of each web service that are provided by the same service provider is a very time consuming process. To deal with this problem, the input dependency and output dependency are also used. However, the output and input dependency analysis is performed just syntactically by comparing the types of input or output parameters of the web services.

Initially, input dependencies, output dependencies and the input/output dependencies between all validated services of each service provider are analyzed. On this basis of this analysis, test cases are generated. A web service may have no input parameter and also may return no value. For each validated web service with at least one input parameter, the dependencies with the other validated web services that are provided by the same service provider are analyzed.

For semantic input/output dependency analysis, similarity between the names of the parameters is found through matching degree calculation. For this calculation phase, the functions of word matching library presented in (Canturk and Senkul, 2011) is used. This matching method is extended from WordNet and it performs both syntactic and semantic matching. It is possible to use any other word matching tool that returns a score for similarity. However, we preferred to use this new matching technique since it has shown to provide promising improvement over similar techniques, especially for semantic matching.

While finding matches, each term in the first pa-

rameter is compared with the all of the terms in second parameter. As the result of this similarity calculation, we obtain a similarity degree array with length of n. However, we desire to get a single value as the similarity degree of whole output term to whole input term. Therefore, the average value of these similarity degrees is calculated for the final similarity value (s1) by using Equation 1.

$$sim(term_1, term_2) = \frac{1}{n}\sum_{j=1}^{n} sim(term_1, term_{2_{word_j}}) \quad (1)$$

We devise also another method to calculate the similarity degree between input and output terms where domain ontology is available. Note this ontology is not for service description, but for describing the domain of the services, such as Car or Movie domain. In this method, the distance between the input term and each ontology term and the distance between each output term and each ontology term are calculated. These distances are used to determine the ontological positions of the terms. By calculating the average of the differences of these distance values with each ontology term, the similarity value (S2) between input term and output term is calculated by using Equation 2.

$$ontsim(term_1, term_2) = 1 - \frac{1}{n}\sum_{k=1}^{n} |sim(term_1, ontTerm_k) - sim(term_2, ontTerm_k)|$$

where n is the number of ontology terms.

The final similarity value with the name of the output term and the web service whose output parameter is analyzed are recorded to input/output dependency list of analyzed input parameter. A sample list of input/output dependency relations of web services is shown in Figure 2.

6 OVERALL TEST SCORE GENERATION

Overall test score of a web service is calculated through its performance in mutation-based and semantic dependency analysis based test generation phases. For all of validated web services, different test cases based on both semantic analysis operation and data mutations are generated.

Firstly, the test case generation based on data mutant analysis is performed. By using each version of mutant value generation mentioned in the previous section, different values are generated for each input parameter. Web service is invoked with these input parameters and return value is obtained. In Mutant Version 0, five different values in the given range are generated for each input parameter. By setting the input parameters to these generated values, the test case is prepared for the web service. The next step is test case execution. In this phase, web services are invoked and the results are checked. If an exception occurs during the execution of the web service request, then the web service is accepted as unsuccessful. This result is recorded to statistics of testing of web service. On the other hand, it is expected that the web services that cause no exception have different return values. Currently, a web service pass the test if it produces different outputs to different inputs. However, the obtained result value will be further analyzed in detail. The similar steps are followed in the other mutant versions. In Mutant Version 1, 2, 3, 4, 5 and 6, just one test case is generated. To test the web service successfully, it is expected that no exception is occurred. In Mutant Version 5 and 6, the possibility of being failed is higher than the other versions because the web services might not handle the values in these ranges. As mentioned for Mutant Version 0, the return values are analyzed in detail and the test results are recorded.

For mutant version in which the values of input parameters from ontology instances are obtained, one test case is generated in which the values of input parameters from WordNet instances are obtained. As the last mutation version, input parameter switching is performed. The test cases that are generated in mutant version 0 and that provide successful test result are used again in switching version, if the parameters are suitable.

As the second phase, test case generation based on dependency analysis is performed. If the web service has no input parameter, it is invoked without generating any input value and the invocation result is analyzed. If it has at least one input parameter, the value set is generated for each input parameter by performing the following steps.

Firstly, the dependency list is analyzed. If it has no dependency with the return parameters of the other web services, considering its type, a random value is generated by using version 0 in the mutant-based method. Otherwise, by starting with the first web service in the dependency list, the test cases of web services is analyzed to get the appropriate value from its return value. The dependency relation list is sorted by dependency degree, therefore the analysis operation is started with the first one. If the first web service in the list has no successful test case, it is not used for

URLID	Method1	Method2	InputParamName	OutputTerm	DependencyDegree
385	getAirportsInCity	getAirports	city	city	1
385	getAirwaysInFlightLine	getFlightLines	arrival	arrivalAirport	1
385	getAirwaysInFlightLine	getFlightLinesFrom	arrival	arrivalAirport	1
385	getAirwaysInFlightLine	getFlightLineOfAirline	arrival	arrivalAirport	1
385	getAirwaysInFlightLine	getFlightLinesTo	arrival	arrivalAirport	1
385	getFlightLinesFrom	getFlightLineOfAirline	takeoff	departureAirport	0.9728947877883...
385	getAirwaysInFlightLine	getFlightLinesTo	takeoff	departureAirport	0.9728947877883...
385	getAirwaysInFlightLine	getFlightLineOfAirline	takeoff	departureAirport	0.9728947877883...
385	getFlightLinesFrom	getFlightLinesTo	takeoff	departureAirport	0.9728947877883...
385	getFlightLinesFrom	getFlightLines	takeoff	departureAirport	0.9728947877883...
385	getAirwaysInFlightLine	getFlightLines	takeoff	departureAirport	0.9728947877883...
385	getFlightLinesTo	getFlightLines	coming	arrivalAirport	0.9676315784454...
385	getFlightLinesTo	getFlightLineOfAirline	coming	arrivalAirport	0.9676315784454...
385	getFlightLinesTo	getFlightLinesFrom	coming	arrivalAirport	0.9676315784454...
385	getFlightLineOfAirline	getFlightLines	airline	airWay	0.9199999570846...
385	getFlightLineOfAirline	getFlightLinesFrom	airline	airWay	0.9199999570846...
385	getFlightLineOfAirline	getFlightLinesTo	airline	airWay	0.9199999570846...
385	getFlightLinesTo	getFlightLineOfAirline	destination	airWay	0.7549999952316...
385	getFlightLinesTo	getFlightLines	destination	airWay	0.7549999952316...
385	getFlightLinesTo	getFlightLinesFrom	destination	airWay	0.7549999952316...
385	getAirwaysInFlightLineByID	getAirports	arrivalAerodromeID	airportID	0.7126315832138...
385	getAirwaysInFlightLineByID	getAirports	takeoffAerodromeID	airportID	0.6963157653808...
385	getFlightLinesTo	getAirports	destination	airportName	0.6857894659042...

Figure 2: Input/Output Dependency Relations of Web Services.

generation and the analysis is continued with the next web service in the dependency list.

7 EXPERIMENTAL EVALUATION

7.1 Analysis Method for Test Results

As we described in previous sections, web service invocation results are obtained by calling a web service by using the input parameters that are generated with data mutation and semantic dependency analysis based method. The test results are generated by examining the invocation results in two dimensions. The first one is based on checking the execution results. An exception can occur during the execution of the web service request. In this case, the web service is accepted as unsuccessful. On the other hand, if it works properly and causes no exception, it is accepted as successful.

The second dimension is based on checking whether the service gives different outputs for different input values. If a service always returns the same value in spite of different input parameters, it is considered as a test service or this service has no implementation. If the same output value is obtained after all successful execution of test cases, the web services is failed for this point.

The other checking is for the input parameter switching method. It is expected that different return values are obtained after the executions of the test cases that use the switching method and the source test cases whose input parameters are switched to generate the new test case. For each test case using input parameter switching method, if the return values of itself and source test case are different, it is accepted as successful in this checking process. Otherwise it fails the test.

By using the obtained results of the web services for two dimensions mentioned above, the final test result is calculated. In testing of a web service, its successful invocation is more important than its returning different values. The web services that throw exception when they are invoked are never preferred by the service consumers. Therefore, the overall success result of the test cases is more weighted than having different outputs for different input values. Therefore, a weighted average is applied, where the weight of first dimension is set to 0.7, under empirical evaluation.

7.2 Experiments with Synthesized Web Services

In this set of experiments, we created web service providers that include various services, whose behaviours are already known. These web services belong to *Aviation* and *Car* domains. For Aviation domain, the provided web services are used to get the information about flights in Turkey.

For each web service, the proposed approach is applied and the test results are obtained after the testing process. As we mentioned in previous sections, different input values are generated for each parameter and with these generated input parameters, web services are invoked. After the invocation, the return values of web services are analyzed.

Table 1: MSE on Synthesized Data Set.

Web Service Name	**Expected(T)**	**Estimated(E)**	**$(E-T)^2$**
sayHello	0.5	0.7	0.04
getAirways	1	0.7	0.09
getAirports	1	0.7	0.09
getFlightLines	1	0.7	0.09
getAirportsInCity	1	1	0
getFlightLinesTo	1	1	0
getFlightLinesFrom	1	1	0
getAirwaysInFlightLine	1	0.85	0.0225
getAirwaysInFlightLineByID	1	0.94	0.0036
getFlightLineOfAirline	1	1	0
isRouteBidirectional	0	0	0
getValue	0.5	0.7	0.04
getAllCities	1	0.7	0.09
getAllCarModels	1	0.7	0.09
getCostumers	1	0.7	0.09
getAvailableCarsInTown	1	0.85	0.0225
getCarBookingsByTownId	1	1	0
getCarBookingsByTownName	0	0.63	0.3969
getCarBookingsByDate	1	0.7	0.09
getCarBookingsOfPerson	1	1	0
insertNewCarBookingItem	1	0.97	0.0009
MSE	0.2346		

Table 2: Statistical Information on Web services in Car Domain.

Total number of web service providers	904
Total number of validated web services	5713
Total number of input parameters	15284
Total number of string input parameters	9966
Total number of string input parameters whose value can be assigned from WordNet instances	1426
Average number of services per service provider	6.3
Average number of input parameter per web service	2.7

In order to obtain the accuracy of our proposed algorithm, we calculate the root mean square error (RMSE). Accuracy results for the syntactically generated web services and average accuracy are presented in Table 1. The proposed testing method predicts the reliability of this set of web services with 0.2346 accuracy error. Since the similar studies in the literature do not produce a comparable overall test score, it is not possible to make a direct comparison with the literature. However, this error value is promising for the usability and effectiveness of the proposed method.

7.3 Experiments with Real Web Services

In addition to experiments with synthetic data set, the proposed method is tested on a set of real web services collected in *Car, Aviation, Film* and *Sports* domains. In this set of experiments, it is not possible to obtain a success rate, as the services are not annotated. The results rather provides an approximate picture of the current situation for the reliability of the published services. The web services are collected through web search given the domain name as the keyword. The number of the web services that are provided by the same provider, and the number and type of the parameters have a high variation in this collection. Due to space limitation, we present the results for web services in *Car* domain.

Some of the statistical information for the web services collected in Car domain is given in Table 2. In this collection, out of all web services, 349 services have the overall test score of 1 hence they successfully passed the test cases. The overall test score of 2545 web services is 0, which means that they are totally failed in the test process. 2501 web services are successfully invoked in all test cases of data mutation-based method. 2649 web services are successful in all test cases that are generated by using semantic dependency-based method. 2474 web services are successful in both of two methods. Number of web services that return same value under switching of input values is 2453. Sample results for this set of web services is given in Figure 3.

8 CONCLUSION

In this work, we proposed a method to test discovered web services and to generate a test scores automatically. To realize the proposed method, an application is designed and a graphical user interface is also presented to test the web services automatically through running the method step by step for a given web service.

In our approach, data mutation based and seman-

ID	ServiceDescriptionID	DateTime	FinalResult	SuccessResult	DifferenceResult	isTestMethod	DifferenceOfSwitchVersion
18	226	25.10.2011 03:49:05	1	1	1	1	NULL
19	227	25.10.2011 03:49:05	1	1	1	1	NULL
20	237	25.10.2011 03:50:38	0.042	0.06	0	0	NULL
21	238	25.10.2011 03:50:38	1	1	1	1	NULL
22	239	25.10.2011 03:50:38	0.7	1	0	0	0
23	240	25.10.2011 03:50:38	0.7	1	0	0	0
24	241	25.10.2011 03:50:38	0.7	1	0	0	0
25	242	25.10.2011 03:50:38	0.85	1	0.5	1	0
26	243	25.10.2011 03:50:38	0.88	1	0.6	1	0.2
27	244	25.10.2011 03:50:38	0.7	1	0	0	0
28	245	25.10.2011 03:50:38	0.85	1	0.5	1	0
29	246	25.10.2011 03:50:38	0.7	1	0	0	0
30	247	25.10.2011 03:50:38	0.8127272...	0.8181818181...	0.8	1	0.6
31	248	25.10.2011 03:50:38	0.7	1	0	0	0
32	249	25.10.2011 03:50:38	0.7	1	0	0	NULL
33	250	25.10.2011 03:50:38	0.7	1	0	0	NULL
34	251	25.10.2011 03:50:38	0.7	1	0	0	NULL
35	252	25.10.2011 03:50:38	0.7	1	0	0	NULL
36	253	25.10.2011 03:50:38	0.91	1	0.7	1	0.4
37	254	25.10.2011 03:51:40	0.042	0.06	0	0	NULL
38	255	25.10.2011 03:51:40	0.042	0.06	0	0	NULL
39	256	25.10.2011 03:51:40	0.7	1	0	0	0
40	257	25.10.2011 03:51:40	0.042	0.06	0	0	NULL
41	258	25.10.2011 03:51:40	0.7	1	0	0	0
42	259	25.10.2011 03:53:06	0.042	0.06	0	0	NULL
43	260	25.10.2011 03:53:06	0.7	1	0	0	NULL
44	261	25.10.2011 03:53:06	0.7	1	0	0	NULL

Figure 3: Sample test results for Web Services in Car Domain.

tic dependency based input data generation methods are used. In data mutation, web services are tested by using random and specific values in different ranges. Values in the specified ranges are generated according to the input type. Different data mutation groups are constructed to generate values for input parameters. Since the internal working mechanism of web services are not available in practice, and for most of the time, the exact respond is not known as well, we prefer to use simple, yet as much as possible techniques for testing.

To be able to evaluate the success of the proposed method, we generated a synthetic data set whose behaviour is known. In this data set, error value is 0.23 under RMSE. It is not possible to make a direct comparison with similar studies since they do not generate such an overall test score. However, this error value is promising for the applicability of the approach. In addition to synthetic data set, we tested the real web services. In this evaluation, it is not possible to provide a success rate as these services are not annotated. This evaluation show that the number of web services that can pass all test cases is about 20%.

REFERENCES

AbuJarour, M. and Oergel, S. (2011). Automatic sampling of web services. In *Proc. IEEE Int. Web Services Conf. (ICWS)*, pages 291–298.

Bai, X., Dong, W., Tsai, W. T., and Chen, Y. (2005). WSDL-based automatic test case generation for web services testing. In *Proc. of SOSE*, pages 207–212.

Canturk, D. and Senkul, P. (2011). Semantic annotation of web services with lexicon-based alignment. In *Proc. of IEEE 7th World Congress on Services (SERVICES)*, page 355362.

Dranidis, D., Kourtesis, D., and Ramollari, E. (2007). Formal verification of web service behavioral conformance through testing. *Annals of Mathematics, Computing and Teleinformatics*, 1:36–43.

Martin, E., Basu, S., and Xie, T. (2006). Automated robustness testing of web services. In *Proc. of the 4th SOAWS*.

Siblini, R. and Mansour, N. (2005). Testing web services. In *Proc. 3rd ACS/IEEE Int. Computer Systems and Applications Conf.*

Wang, Y., Bai, X., Li, J., and Huang, R. (2007). Ontology-based test case generation for testing web services. In *JProc. of ISADS*, page 4350.

Evaluating Web Service QoE by Learning Logic Networks

Natalia Kushik[1,2], Nina Yevtushenko[1], Ana Cavalli[2], Wissam Mallouli[3] and Jeevan Pokhrel[2,3]
[1]*Department of Information Technologies, Tomsk State University, Tomsk, Russia*
[2]*Department of Network Software, TELECOM SudParis, Evry, France*
[3]*Montimage, Paris, France*
ngkushik@gmail.com, ninayevtushenko@yahoo.com, ana.cavalli@it-sudparis.eu,
wissam.mallouli@montimage.com, jeevanpokhrel@gmail.com

Keywords: Web Service, Quality, QoE, Logic Network/Circuit.

Abstract: This paper is devoted to the problem of evaluating the quality of experience (QoE) for a given web service based on the values of service parameters (for instance, QoS indicators). Different self-learning algorithms can be used to reach this purpose. In this paper, we propose to use self-learning logic networks, called also circuits, for evaluating the QoE of web services, since modern software tools can efficiently deal with very large logic networks. As usual, for machine learning techniques, statistics are used to design the initial circuit that accepts service parameter values as inputs and produces the QoE value as an output. The circuit is self-adaptive, i.e., when a new end-user provides a feedback of the service satisfaction the circuit is resynthesized in order to behave properly (if needed). Such resynthesis (circuit learning) can be efficiently performed using a number of tools for logic synthesis and verification.

1 INTRODUCTION

As the number of web services increases quickly, more methods and tools are needed to evaluate their quality. Moreover, there exist a number of service repositories (Directory of Public SOAP Web Services; Online community for the Universal Description, Discovery, and Integration OASIS Standard). Thus, a question arises of choosing the best service among the set of services with the same functionality. In order to effectively perform such a service selection, a 'selection criteria' need to be formalized. As usual, such formalization is based on the service quality that is defined with respect to different metrics or indicators.

One of most common metrics to evaluate a web service quality is a QoS (Quality of Service). The QoS is usually represented as a set of attributes or parameters of this service (Khirman and Henriksen, 2002; Al-Masri and Mahmoud; Kim et al., 2008; Morais and Cavalli, 2012). Those parameters could be the service response delay, packet loss rate, reliability, etc. Such parameters are rather objective and the value of each parameter can be automatically measured relying on specific tools. Nevertheless, the main objective of every web service is the satisfaction of end-users and thus, other subjective parameters need to be taken into account when estimating the QoE (Quality of Experience).

As the QoE is used to measure the end-user satisfaction, the problem of its evaluation is much more complicated than the QoS evaluating problem. In other words, in order to evaluate QoE it is necessary to 'guess' how much an end-user would like or dislike the web-service. Thus, the problem arises of modeling the user perception, that is nowadays one of most popular problems in the artificial intelligent area where self-adaptive models are used for this purpose. A self-adaptive model usually accepts service parameter values as inputs and returns the QoE value as an output. If a model behaves in a wrong way for some newly emerged input/output pairs the model can be trained by itself or by an external 'teacher' that could be a service provider. The most popular of these self-adaptive models are discussed in the section on related work.

In this paper, we are focusing on a novel self-adaptive model approach for evaluating and 'predicting' the QoE value of web services. In fact, we extend an approach based on decision tree learning (Mitchell, 1997) to logic network (circuit) learning. As a decision tree can be represented as a disjunctive form where each conjunction represents a path to the leaf, we are looking forward to a

scalable representation and manipulation of such Boolean formula. One of possible ways to effectively represent such formula is to use a logic circuit. Similar to a decision tree, the initial logic network is derived based on statistics that are gathered from experts, developers and those end-users who agreed to provide a feedback about the service quality. The circuit accepts the service parameter values encoded as Boolean vectors and outputs the Boolean vector that corresponds to the encoded QoE value. The circuit is a self-learning machine, i.e., when new statistical data appear the circuit is checked to have the corresponding behavior and if the behavior does not correspond to newly emerged data the circuit is resynthesized. Such resynthesis can be efficiently performed using various tools. In this paper, we use the ABC (Berkeley Logic Synthesis and Verification Group, ABC) tool for logic synthesis and verification to train the corresponding circuit in a scalable way. The main contribution of this paper is the proposal of a machine learning technique for automatic evaluating QoE of web services, in which a machine is represented by a logic circuit and a learning operator corresponds to the circuit resynthesis.

An iterative approach is known to be efficient for logic circuit optimization. Given a circuit under optimization, the circuit is incrementally improved at each optimization step. In this paper, we consider a special optimization criterion, namely, the circuit QoE 'prediction' ability, and illustrate that an iterative learning of the circuit under construction is promising.

The rest of the paper is organized as follows. Section 2 contains the preliminaries and basic concepts related to logic network paradigm, while a state of the art of using self-adaptive models for the QoE evaluation is given in Section 3. The approach on using logic networks for the QoE evaluation is presented in Section 4. Section 5 contains an algorithm for the corresponding logic network learning procedure and a brief discussion of a proposed approach. Section 6 contains preliminary experimental results while Section 7 concludes the paper.

2 PRELIMINARIES

A *web service* can be described as a software system designed to support interoperable machine-to-machine interaction over a network (Booth et al., 2004). The Quality of a web Service (QoS) is defined as a vector with components which are values of given attributes (parameters), such as time delay, packet loss rate, etc. The QoS is a metrics that represents some objective service parameter values that can be effectively measured, for example, based on the traffic analysis (Khirman and Henriksen, 2002). The QoE metrics is more complicated, since it represents a user satisfaction (Winckler et al., 2013) with a web service. Generally, the QoE also relies on different service parameters that are rather subjective than objective, those can be service design, comfort of proposed solutions, etc. In spite of the fact that the QoE is more difficult to evaluate, this metrics is more close to adequate description of the service quality, since the main purpose of each web service is to satisfy an end-user. In other words, the algorithm for the QoE evaluation has to be adapted to a human's brain in order to 'predict' what a user likes/dislikes. That is the reason why different self-adaptive models and algorithms are now used for this purpose. The advantage of a self-adaptive model is that it can be learnt or trained by a 'teacher' or by itself according to the feedback from people who use the service. As usual, an initial model/machine is derived based on some statistics that contain a number of user opinions about the service. Afterwards, the model can 'predict' a user satisfaction of the service for given values of service parameters. The more statistics are gathered the better is such 'prediction'. Moreover, as a model is self-adaptive, when new statistical data appear for which the model does not behave in an appropriate way, the model is adjusted to this new data and this process is the model training.

Given a self-adaptive model, different service parameters might be considered. One of most often case is to use QoS parameters, as those values can be measured automatically. On the other hand, for the QoE prediction the values of more subjective parameters, for instance, related to the business or trust quality should be considered as well.

In this paper, we are focusing on logic networks to be learnt for the QoE 'prediction' and we further briefly sketch the necessary definitions related to logic synthesis.

Definition 1. A *logic network* (circuit) consists of logic gates. Each logic gate has input (-s) and a single output. Outputs of some gates are connected to inputs of the others. The inputs of some gates that are not connected to any other gate output are claimed to be primary inputs while the outputs of some gates are claimed as primary outputs. In this paper, we consider combinational circuits, i.e., feedback-free circuits which have no latches.

Each gate implements a Boolean function. Most

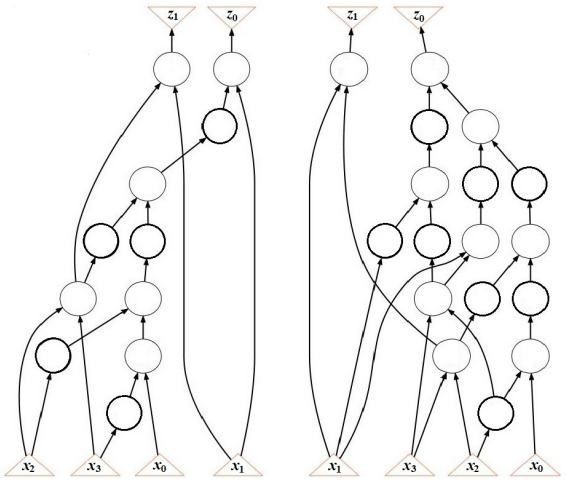


Figure 1: a) A circuit S, b) A circuit S'.

common 2-input gates are AND/OR/XOR/NAND/NOR/XNOR that implement conjunction/disjunction/xor and their inversions. There are also 1-input gates such as NOT/BUFF that implement the inversion and the equality function, correspondingly.

As an example, consider a combinational circuit in Fig. 1a with a set $X = \{x_0, x_1, x_2, x_3\}$ of inputs, a set $Z = \{z_0, z_1\}$ of outputs and 11 AND and NOT gates. Hereafter, we assume that NOT nodes are taken in bold.

Definition 2. By definition, a logic circuit *implements* or *represents* a system of Boolean functions. A circuit accepts a Boolean vector as an input and produces the Boolean vector as an output according to the corresponding system of Boolean functions. Each logic circuit can be described by a Look-up-Table (LUT). A LUT contains a set of input/output pairs of a given circuit: if for the input **i** the circuit produces an output **o**, then the pair **i**/**o** is included into the LUT (see, for example, Table 2).

A LUT can be used as the specification when deriving a logic network that implements the system of Boolean functions, and there exist a number of methods how to synthesize a logic network that implements a given system of functions.

If this system is completely specified then 2-level network can be synthesized based on corresponding Disjunctive Normal Form (DNF) or Sum of Products (SoP) (McCluskey, 1965); otherwise, if the system of Boolean functions is only partially specified, then other methods are used for logic synthesis (see, for example, Kuehlmann, 2003). In this case, each input pattern where the circuit behavior is not specified is usually treated as a 'Don't_Care' pattern, i.e., a circuit under design can produce any output to this input. Correspondingly, the circuit behavior is specified for such 'Don't_Care' patterns according to the designer's needs. It can be guarded by optimal criteria of a circuit under design, such as number of logic gates, time needed to produce an output, etc. In this paper, we are interested in deriving a circuit that models the QoE evaluation of web services, thus, one of optimal criteria could be the accuracy of the circuit prediction.

3 RELATED WORK

As mentioned above, the main priority of each web service is to satisfy an end-user, thus a service provider is trying to 'guess' what the end-user would like and what will be disliked. Therefore, the problem arises of modeling user perception that is nowadays one of most popular problems in the artificial intelligent area. A number of continuous and discrete models are developed for this purpose but nevertheless, no one can prove or disprove the accuracy of these models. Since a human mind is known to be non-deterministic, just now it is not possible to get an algorithm of the human behavior. That is the reason why researchers are focused on the study of self-adaptive models. Several self-adaptive models have been experimentally proven to behave similarly to a human being under certain conditions. In this paper, we do not discuss all the self-adaptive models that are developed in this artificial intelligent research area, instead we focus on those self-adaptive models that can be used to evaluate a user satisfaction of a given web service.

Almost all self-adapting models for evaluating the quality of web services rely on pre/post conditions that can be expressed as IF-THEN operators. In order to analyze a set of such conditions, machine learning algorithms can be of a big help. One of such algorithms is based on a Decision Tree (Mitchell, 1997; Pokhrel et al., 2013) that can be usually represented as a Boolean formula, i.e., a disjunction of conjunctions (cubes). Nevertheless, the size of a decision tree grows exponentially w.r.t. the number of parameters to be verified and thus, other more compact learning models that can be derived on the IF-THEN conditions should be considered.

One of popular models is a model of an Artificial Neural Network (see, for example, Ahmed et al., 2012; Al-Masri and Mahmoud Qusay, 2009) that accepts inputs and produces outputs depending on hidden nodes, their weights and connections. Neural

networks are widely used in various applications, such as modeling company marketing, oil flows, etc. Meanwhile, a number of methods for evaluating the QoE of web services are based on various kinds of neural networks. This network can be trained by an external 'teacher' or by itself to produce more precise results, i.e., it can be somehow adapted to the statistics being gathered from end-users of the service.

Another interesting model that can be used for the QoE evaluation is a fuzzy logic formula (Zadeh, 1965) where the operands do not have precise values but belong to some intervals which can be rather informal. In this case, a fuzzy logic formula represents a degree of a user satisfaction with the service. One may turn, for example, to papers (Lin et al., 2005; Torres et al., 2011) to see how fuzzy logic formulae can be used for evaluating QoE of web services.

Neural network and fuzzy logic formulae are experimentally proven to provide a reasonably accurate prediction of the user satisfaction. Nevertheless, when using these models for the QoE evaluation of web services as well as for the learning process a researcher faces a number of computational problems. The reason is that both models deal with continuous variables and thus, when evaluating the QoE floating-point numbers with a lot of decimal points are involved. As the discrete model, a decision tree does not have this drawback. On the other hand, a decision tree with the high prediction ability requires exponential memory and correspondingly needs a scalable representation. Such scalable representations for Boolean formulae include logic circuits, and in this paper, the Boolean formula corresponded to a decision tree is represented by a logic network that can 'predict' the QoE value even for patterns that are not specified in the provided statistics.

4 USING LOGIC NETWORKS FOR QOE PREDICTION

In this section, an approach for automatic evaluation/'prediction' of the QoE value for web services is proposed. Since the approach relies on logic circuit learning we first discuss how this (initial) circuit can be derived and then turn to a learning procedure itself.

Consider a web service W and a collection of parameters p_1, p_2, …, p_k that are used for the QoE evaluation. For the sake of scalability, we consider each parameter p_i value as a nonnegative (unsigned) integer, bounded by the maximal value M_{p_i}, $p_i\_value \in \{0, \ldots, M_{p_i} - 1\}$. In order to evaluate the QoE of the service W, written $QoE(W)$, a logic circuit S is derived. Inputs of the circuit S correspond to service parameters p_1, p_2, …, p_k which values are encoded as Boolean vectors of length $]\log_2 M_{p_i}[$, where $]t[$ denotes the minimal integer that is not less than t; thus, the number of primary inputs of S equals $\sum_{i=1}^{k}]\log_2 M_{p_i}[$. The circuit calculates the $QoE(W)$ value for given integers p_1_value, p_2_value, …, p_k_value and $QoE(W)$ that is bounded by the maximal value M_{QoE}, $QoE(W) \in \{0, \ldots, M_{QoE} - 1\}$, that is also represented as a Boolean vector. The length of a corresponding Boolean vector is $]\log_2 M_{QoE}[$, as $]\log_2 t[$ bits are needed to represent an unsigned integer t. Therefore, the circuit S has $\sum_{i=1}^{k}]\log_2 M_{p_i}[$ primary inputs and $]\log_2 M_{QoE}[$ primary outputs.

4.1 Using Statistics for Circuit Learning

In order to derive the initial circuit S, we use statistics gathered from web service experts, from the automatic evaluation of service parameters and/or from end-users, who have experience of using a web service W. A running example is used for illustrating how such statistics can be utilized for logic synthesis.

Example. Let us consider an electronic service for booking a student dormitory. When entering the web site a student is asked for his personal information such as his/her name and family name and at the second step he/she indicates dates of the desired stay. The booking system checks the availability for the student to stay at the dormitory and outputs the result.

After using such electronic service, a student is asked whether he/she is willing to evaluate its quality and to provide his/her feedback to the service provider. If the student agrees then he/she is asked for evaluating the design quality of the system as well as the speed of it. Thus, two parameters p_1 and p_2 are considered where p_1 represents the service design while p_2 represents its speed. In this toy example, we assume that each parameter is evaluated as an integer between 0 and 3. The overall user satisfaction which value belongs to the same

interval is also asked for the student evaluation. To sum it up, the student is asked three questions:

1. How do you find the design of the electronic service? Please evaluate the design by an integer between 0 and 3 (0 means that the service design is very bad and 3 means that the service design is excellent).

2. How do you find the speed of electronic service interaction? Please evaluate the speed by an integer between 0 and 3.

3. *What is your overall User Satisfaction within this electronic service? Please evaluate your satisfaction by an integer between 0 and 3.*

A student can ignore the above questions but if a student replies then new statistical data are retrieved about the service quality.

Consider three users who have already left their scores. The user *A* evaluates the design quality and the speed of the service by the highest score and thus, the QoE evaluation is 3 (*excellent*). The user *B* gives the same evaluation for speed while he/she does not like the design (the score is 1) that does not influence his/her overall satisfaction that equals 3. The last user *C* that gives a feedback likes the design (the score is 3) while the speed is scored as 0, and the overall satisfaction is 2 (Table 1).

Table 1: Statistics gathered for dormitory booking electronic service.

User identifier	Design score	Speed score	User Satisfaction
A	3	3	3
B	1	3	3
C	3	0	2

The statistics described above are converted into input/output vectors for the circuit *S*. The circuit has two inputs for representing the value of the *design_score*, two inputs for representing the value of the *speed_score* and two outputs for representing the *QoE* (*dormitory_booking*). Based on statistics in Table 1 we derive a LUT (Look-up-Table) for a system of two output functions z_0, z_1, depending on four input variables denoted as x_0, x_1, x_2, x_3, i.e., we partially specify the behavior of the circuit *S* (Table 2). Table 2 describes the system of partially specified Boolean functions and thus, it can be used as the specification for designing a corresponding logic circuit.

Table 2: LUT for the statistics from Table 1.

x_0	x_1	x_2	x_3	z_0	z_1
1	1	1	1	1	1
0	1	1	1	1	1
1	1	0	0	1	0

4.2 Designing a Logic Circuit by ABC

We have derived the circuit *S* for the system of partially specified Boolean functions in Table 2 using the software tool ABC (Berkeley Logic Synthesis and Verification Group, ABC) that is a growing software system for synthesis and verification of binary synchronous sequential logic circuits. ABC provides scalable logic synthesis and optimization based on And-Inverter Graphs (AIGs) and other internal circuit representations. Moreover, ABC includes a number of commands for synthesis, resynthesis, optimization and verification of logic circuits as well as for printing the circuit statistics and parameters. The ABC can be easily downloaded from its official web site and run against a LUT for web service *W* to synthesize a corresponding circuit *S*. We have run the ABC tool against the LUT corresponded to Table 2. For this purpose, we have presented the set on input/output vectors in the PLA format. The resulting circuit *S* with 11 gates is shown in Fig. 1a. This circuit computes the QoE value of the dormitory booking service based on statistics in Table 1. If a service provider wants to 'predict' the overall user experience then the output of the circuit *S* can be automatically calculated for a corresponding input vector.

The logic circuit *S* that is derived based on the provided statistics, evaluates and 'predicts' the QoE value of web service *W* for any values of input parameters. The Don't_Care patterns can be used in a proper way when deriving the initial circuit for the QoE evaluation and below we present an algorithm for deriving the initial circuit *S*.

Algorithm 1 for deriving an initial logic circuit to evaluate the QoE value

Inputs: Service parameters p_1, p_2, …, p_k with nonnegative (unsigned) integer values bounded by M_{p_1}, M_{p_2}, …, M_{p_k}; maximal value of the QoE M_{QoE};

Statistics, i.e., feedbacks from users U_1, …, U_r represented as a list of patterns p_1*_value*, p_2*_value*, …, p_k*_value*, *UserSatisfaction_value*.

Output: a logic circuit *S*

1. Determine the number of primary inputs and primary outputs of *S*:

The number of primary inputs equals $\sum_{i=1}^{k}]\log_2 M_{p_i}[$ while the number of primary outputs equals $]\log_2 M_{QoE}[$.

2. Derive a LUT

2.1 For each user U_i, $i \in \{1, \ldots, r\}$, convert its statistic scores p_1*_value*, p_2*_value*, …, p_k*_value*, *UserSatisfaction_value* into Boolean vectors and add the corresponding line to LUT.

3. Synthesize the circuit S from a partial system of Boolean functions and **Return *S***.

□

Given an integer x, let **B**(x) denote a corresponding Boolean vector for x. Due to steps 2 and 3 of the algorithm, the following statement holds.

Proposition 1. Given a service W and a statistic pattern p_1*_value*, p_2*_value*, …, p_k*_value*, *UserSatisfaction_value*, a circuit S returned by Algorithm 1 produces the output **B**(*UserSatisfaction_value*) when the concatenation **B**(p_1*_value*) · **B**(p_2*_value*) · … · **B**(p_k*_value*) is applied as an input vector.

Once a circuit S is derived, the circuit can be used for evaluating the QoE of the web service. As in this paper we are interested in self-adaptive models, we discuss how such circuit can be modified if the circuit behavior does not match new statistical data.

5 LOGIC NETWORK LEARNING PROCEDURE

A circuit S derived in the previous section has to be self-adaptive, i.e., when a new end-user agrees to leave his/her feedback about the service quality the circuit behavior has to be modeled under a corresponding input **i** and if the result produced by the circuit differs significantly from the expected then the circuit has to be resynthesized. To evaluate the difference between the circuit output and the user satisfaction value we introduce some value τ that represents a confidence interval, i.e., the $QoE(W)$ produced by the circuit S has to belong to the interval [*UserSatisfaction_value* − τ, *UserSatisfaction_value* + τ]. This fact can be considered as an optimization step when constructing a logic network for evaluating/'predicting' QoE of web services. Given a new pattern **i**/**o** where **o** represents the *UserSatisfaction_value* and $QoE(W)$ is an integer that corresponds to the S output for the input **i**, if $|QoE(W) - UserSatisfaction\_value| \leq \tau$ then the circuit S is not resynthesized.

The confidence interval [*UserSatisfaction_value* − τ, *UserSatisfaction_value* + τ] specifies the permissible distance between circuit behavior and new statistical data. If the $QoE(W)$ value produced by the circuit S does not belong to this interval several cases are possible:

1) The input part **i** of a new statistical pattern **i**/**o** where **o** is a Boolean vector for the expected *UserSatisfaction_value*, was not in the initial LUT (not described in the previous statistics); and thus, the circuit S was not correctly synthesized for a Don't_Care input **i**. In this case, the new input pattern **i**/**o** is added to the LUT for the circuit and a circuit is resynthesized.

2) The output of the circuit S for the input **i** was defined in the initial LUT but the output is not within the confidence interval. If this mismatching is caused by an expert error that provides initial statistics, we consider an end-user as an expert and resynthesize the circuit w.r.t. the new output for the input **i**.

3) The output of the circuit S for the input **i** is not within the confidence interval and this mismatching is due to different end-user opinions. In other words, for the same service parameter values, two different users specify the user satisfaction values with the difference greater than the constant τ. In this case, a problem arises of solving a conflict of user opinions. In this paper, we do not go deep into solving this problem, since the probability of such situation is very low when the number of service parameters is large enough. Thus, we again rely on the last end-user opinion and resynthesize the circuit in a corresponding way.

Algorithm 2 for learning the logic circuit that evaluates/ 'predicts' the QoE value for web service

Inputs: QoE parameters p_1, p_2, …, p_k with nonnegative values bounded by M_{p_1}, M_{p_2}, …, M_{p_k}; maximal value of the QoE M_{QoE};

The circuit S that evaluates the QoE value for web service W;

A new user feedback p_1*_value*, p_2*_value*, …, p_k*_value*, *UserSatisfaction_value*;

Maximal difference τ for corresponding confidence interval.

Output: a modified logic circuit S

1. Integers p_1*_value*, p_2*_value*, …, p_k*_value*, *UserSatisfaction_value* are converted into Boolean vectors **v_p1, v_p2, …, v_pk, v_us**.

2. The output $QoE(W)$ of the circuit S is computed for the input **v_p1, v_p2, …, v_pk**.

3. If $|\ QoE(W) - UserSatisfaction\_value\ | > \tau$ then

3.1 *If* the line **v_p1, v_p2, …, v_pk** is specified as input in the LUT, then change the corresponding output into **v_us,**

Otherwise

Add the new line **v_p1, v_p2, …, v_pk, v_us** to the LUT.

3.2 Synthesize the new circuit S'; assign $S = S'$ and **Return** S.

□

Similar to Proposition 1, the following statement holds.

Proposition 2. Given a service W, a statistic pattern p_1_value, p_2_value, …, p_k_value, *UserSatisfaction_value* and the length τ of a confidence interval, a circuit S returned by Algorithm 2 produces the output $\mathbf{B}(x)$, such that $| x - UserSatisfaction\_value | \leq \tau$ when the concatenation $\mathbf{B}(p_1\_value) \cdot \mathbf{B}(p_2\_value) \cdot \ldots \cdot \mathbf{B}(p_k\_value)$ is applied as an input vector.

In the running **example**, consider another user D that agrees to provide a feedback about using the electronic service of the dormitory booking system, and an output of the circuit S is computed according to his/her feedback. If S provides the output that does not belong to the user D confidence interval with $\tau = 1$, then the circuit should be resynthesized. Suppose the user D has scored the service design by 2 and its speed by 1 while the overall user satisfaction he/she has evaluated as 2, thus the QoE value is 2. These scores converted into Boolean vectors are equal to (1001)/(10). The circuit S in Fig. 1a outputs (00) which corresponds to 0 score when an input (1001) is applied, thus it has to be resynthesized taking into account this new input/output pair. We have performed such resynthesis using ABC and obtained another circuit S' with 17 gates (Fig. 1b).

As soon as necessary statistics is gathered, Algorithm 1 can be applied to synthesize the initial circuit S. In this case, the initial LUT contains r lines where r is number of feedback patterns. As it is well known, the problem of gathering such statistics as well as the problem of possible future circuit resynthesis are hard computational problems. Thus, we need appropriate heuristics in order to minimize the number of 'hard' operations. One of possible solutions could be to specify some Don't_Care patterns before applying Algorithm 1 according to the value τ related to the output confidence interval. In other words, a confidence interval can be considered for inputs as well, and all patterns that are close enough (with respect to this input confidence interval) can correspond to the same output. The latter means that for each input **v_p1, v_p2, …, v_pk,** a set of 'close' inputs can be determined and every input of this set is added to the LUT with the same **v_us** output. The procedure can be simple enough if such confidence interval is represented by a corresponding conjunction (cube) but this issue needs more research. Additional experimental research is also needed to evaluate the optimal length of confidence intervals for different services, since the length of confidence intervals significantly influences both, the accuracy of predictions and the computational complexity contradicting each other. The bigger is the τ value the less resynthesis steps are needed, along with reducing the accuracy of circuit predictions. Studying optimal values for the length of input and output confidence intervals remains one of challenging topics for future work.

6 EXPERIMENTAL RESULTS

We applied the logic circuit learning approach for evaluating the quality of media services using the statistics gathered by an innovative French company Montimage (more details about the company at montimage.com). In order to collect the statistical data for the QoE estimation of media services, Montimage researchers have conducted a subjective test with 25 end-user participants. This test involves a video service QoE evaluation based on network perturbations. Different video clips have been presented to the participants and according to the perceived video impairment, participants have evaluated each video clips quality as an integer from 1 to 5. The test had twenty seconds videos of different types (sports, movie, animation, and interview). To perform experiments, video clips have been streamed from a server to a client and corresponding perturbation has been added into an emulated network. *Packet loss* and *jitter* have been selected for perturbation where *jitter* corresponds to the variation in video packet delays and *packet_loss* is the percentage of packets being lost during video transmitting. In the subjective test, video clips were randomly provided for participants which had five minutes break after watching 20 video clips. Given statistics gathered as described above, we have randomly extracted a sample of 17 user statistical data (Low number of users since we are only dealing a proof of concept experimentation). In this sample, the *jitter* ranges between 2 ms and 20 ms while the *packet_loss* ranges from 0,02% to 0,25%. The latter means, five circuit inputs can be used to represent each parameter scores. To represent the QoE value, the corresponding circuit under construction has three outputs. We also mention that during the experiment we considered the output confidence

difference $\tau = 1$. Moreover, in this sample we also consider the corresponding confidence interval for each input with the same $\tau = 1$ difference.

The current sample has been divided into three parts. At the first step, we have used 13 patterns to derive the initial circuit. Each pattern has been converted into a Boolean vector of length 13. We have run the ABC system against the corresponding system of partial Boolean functions represented in the PLA format. The obtained circuit *S* contains 121 gates while the maximal length of the path from primary inputs to primary outputs equals 22.

Following the given statistics, we have randomly chosen two more input patterns and simulated the behavior of the circuit *S* for new patterns. The simulation results coincide with the given statistical data for the user satisfaction with respect to the output confidence interval. Therefore, the circuit was not resynthesized.

Furthermore, we have added two additional patterns to the corresponding PLA file. We have simulated the circuit behavior on corresponding inputs and for one input the difference between expected and produced outputs has reached two. Correspondingly, the circuit has been resynthesized according to the new statistical data. Finally we have obtained the combinational circuit with 154 AIG nodes and with the maximal length from primary inputs to primary outputs being equal 29.

Preliminary experimental results show that updating the circuit with respect to new statistical data does not significantly increase the length of the path from primary inputs to primary outputs that can influence the speed of simulation procedure. In fact, given statistical data for the media service, the ABC system has produced the logic synthesis results almost immediately. This allows to claim that our proposed approach is promising, as the simulation procedure for the QoE calculation can be efficiently performed.

Further experimental research is needed to estimate the efficiency of this approach for different types of web services and their parameters.

7 CONCLUSIONS

In this paper, we have discussed different self-adaptive models for automatic QoE evaluation/'prediction' of web services. Extending the decision tree approach we have proposed a scalable approach for learning corresponding Boolean formulae using logic circuits for their representation. Two algorithms are proposed in the paper, namely an algorithm for deriving an initial logic network based on statistics being gathered, and an algorithm for logic network learning based on new statistical data. As the future work, we plan to perform more experiments with different web services.

ACKNOWLEDGEMENTS

The work is partially supported by ANR PIMI project (France) and by RFBR grant № 14-08-31640 мол_а (Russia).

REFERENCES

Ahmed, S., Begum, M., Hasan Siddiqui, F., Abul Kashem, M., 2012. Dynamic Web Service Discovery Model Basedon Artificial Neural Network with QoS Support. *International Journal of Scientific & Engineering Research Volume 3, Issue 3*, *pp. 1-7*.

Al-Masri, E., Mahmoud, Q. H. Quality of web services dataset. *url: http://www.uoguelph.ca/~qmahmoud/qws/*.

Al-Masri, E., Mahmoud Qusay, H., 2009. Discovering the Best Web Service: A Neural Network-based Solution. *SMC 2009, pp. 4250-4255*.

Berkeley Logic Synthesis and Verification Group, ABC: A System for Sequential Synthesis and Verification, *url: http://www.eecs.berkeley.edu/~alanmi/abc/*.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., Orchard, D., 2004. Web services architecture. *W3C Working Group Note, W3C Technical Reports and Publications, url: http://www.w3.org/TR/ws-arch/*.

Directory of Public SOAP Web Services, *url: http://www.service-repository.com*.

Khirman, S., Henriksen, P., 2002. Relationship between Quality-of-Service and Quality-of-Experience for public Internet service. In *Proc.of PAM 2002*.

Kim, H. J., Lee, D. H., Lee, J. M., Lee, K. H., Lyu, W., Choi S.G., 2008. The QoE evaluation method through the QoS-QoE correlation model. In *Proc. of NCM 08, vol. 2, pp. 719-725*.

Kuehlmann, A. (ed.), 2003. *The Best of ICCAD. 20 Years of Excellence in Computer-Aided Design*. Kluwer Academic Publishers.

Lin, M., Xie, J., Guo, H., Wang, H., 2005. Solving QoS-driven web service dynamic composition as fuzzy constraint satisfaction. *EEE 2005, pp. 9-14.*

McCluskey, E. J., 1965. *Introduction to the theory of switching circuits*. McGraw-Hill Book Company, NY.

Mitchell, T.M., 1997. *Machine learning*. McGraw Hill series in computer science, McGraw-Hill.

Morais, A., Cavalli, A., 2012. Deliverable D2.1 – State of the art of SQM/CEM technology, tools, and

standartization. *IPNQSYS European project, url: http://projects.celtic-initiative.org/ipnqsis/IPNQSIS-D21.pdf*.

Online community for the Universal Description, Discovery, and Integration OASIS Standard, *url: http://uddi.xml.org*.

Pokhrel, J., Mallouli, W., Montes de Oca, E., 2013. QoE Prediction and Self-Learning Mechanisms. *Technical report on the PIMI Project*.

Torres, R., Astudillo, H., Salas, R., 2011. Self-Adaptive Fuzzy QoS-Driven Web Service Discovery. In *IEEE SCC 2011, pp. 64-71*.

Winckler, M. A., Bach, C., Bernhaupt, R., 2013. Identifying user experience dimensions for mobile incident reporting in urban contexts. *IEEE Transactions on Communications, vol. 56, no. 2, pp. 40-82*.

Zadeh, L. A., 1965. Fuzzy sets. *Information and Control, 8 (3), pp. 338–353*.

From Image to Audio Watermarking Using Self-Inverting Permutations

Maria Chroni, Angelos Fylakis and Stavros D. Nikolopoulos
Department of Computer Science and Engineering, University of Ioannina, GR-45110 Ioannina, Greece
{*mchroni, afylakis, stavros*}*@cs.uoi.gr*

Keywords: Watermarking Techniques, Audio Watermarking Algorithms, Self-inverting Permutations, Representations of Permutations, Frequency Domain, Embedding/Extracting Algorithms, Performance Evaluation.

Abstract: The intellectual property infringement in music due to the proliferation of the internet and the ease of creating and distributing identical digital objects has brought watermarking techniques to the forefront of digital rights protection. Towards this direction, a significant number of watermarking techniques have been proposed in recent years in order to create robust and imperceptible audio watermarks. In this work we propose an audio watermarking technique which efficiently and secretly embeds information, or equivalently watermarks, into an audio digital signal. Our technique is based on the main idea of a recently proposed image watermarking technique expanding thus the digital objects that can be efficiently watermarked through the use of self-inverting permutations. More precisely, our audio watermarking technique uses the 1D representation of self-inverting permutations and utilizes marking at specific areas thanks to partial modifications of the audio's Discrete Fourier Transform (DFT); these modifications are made on the magnitude of specific frequency bands. We have evaluated the embedding and extracting algorithms by testing them on various and different in characteristics audio signals that were in WAV format and we have obtained positive results. The algorithms have been developed and tested using the mathematical software package Matlab.

1 INTRODUCTION

Digital watermarking is a technique for protecting the intellectual property of a digital object; the idea is simple: a unique marker or identifier, which is called *watermark*, is embedded into a digital object which may be used to verify its authenticity or the identity of its owners (Grover, 1997; Collberg and Nagra, 2010).

Audio Watermarking. In a copyright protection framework, an audio watermarking technique aims to embed a unique identifier, i.e., the watermark w, into audio's data through mainly the introduction of errors not detectable by human perception. Within the same framework, audio watermarking can be described as the problem of embedding a watermark w in the host signal S producing thus the watermarked audio signal S_w such that w can be reliably located and extracted from S_w even after S_w has been subjected to transformations such as compression, filtering, noise addition, cropping, etc. It is worth noting that, if a watermarked audio signal S_w is copied or transferred through the internet then the watermark w is also carried with the copy into the audio's new location ensuring thus the maintenance of copyright protection.

Recently, a significant number of watermarking techniques have been proposed in the literature in order to create robust and imperceptible audio watermarks. Initial research on audio watermarking dates back to the mid-nineties where Bender et al. (Bender et al., 1996) presented data hiding techniques for audio signals; the first techniques were directly inspired from previous research on image watermarking. A broad range of audio watermarking techniques goes from simple least significant bit (LSB) scheme to the various spread spectrum methods and can be classified according to the domain where the watermarking takes place in frequency, time, and compressed domain (Sharma et al., 2012; Cox et al., 2008; Alsalami and Al-Akaidi, 2003; Hartung and Kutter, 1999).

Motivation. Nowadays, digital audio is a representative sample of internet data that has been subjected to extensive intellectual property violation. Thus, we consider important the development of methods that deter malicious users from claiming others' ownership, motivating thus internet users to feel more safe to publish their work online.

Audio watermarking, in contrast with other techniques, allows audio signals to be available to third internet users but simultaneously carry an "id" that is actually the ownership's proof. This way audio wa-

termarking achieves its target of deterring copy and usage without owner's permission.

Watermarking digital objects such as image, audio, video, text and software enables the proof of ownership on copyrighted objects preventing thus the intellectual property infringement.

Contribution. In this work we present an efficient and easily implemented technique for watermarking audio signals. What is important of the proposed technique is the fact that it suggests a way in which an integer number w can be represented first as a self-inverting permutation π^* and then as an one-dimensional array (or, equivalently, 1D representation). The idea comes from our previous work on image watermarking where the integer watermark number w is represented as a two dimensional array.

More precisely, our proposed algorithm embeds a self-inverting permutation π^* over n elements into an audio signal S by first mapping the elements of π^* into an $n \times n$ matrix A^* and then, based on the information stored in A^*, marking specific areas of audio S in the frequency domain resulting thus the watermarked audio S_w. An efficient algorithm extracts the embedded self-inverting permutation π^* from the watermarked audio S_w by locating the positions of the marks in S_w; it enables us to reconstruct the 1D representation of π^* and, then, obtain the watermark w.

At this point we would like to point out that the primary purpose of the paper is not to fill a gap of the existing audio watermarking methods by proposing a new embedding technique, but to expand the idea used on our previous work and show that it can be efficiently applied for audio watermarking depicting thus the high versatility of the whole concept.

Evaluation. We have evaluated the embedding and extracting algorithms by testing them on various and different in characteristics audio signals that were in WAV format and we had positive results as the watermark was successfully extracted. What is more, the method is open to extensions as the same method might be used with a different marking procedure. Note that, all the algorithms have been developed and tested in MATLAB (Ingle and Proakis, 2010).

2 OUR WATERMARKING TOOLS

In this section we present the structural and algorithmic tools we use towards the watermarking of an audio signal. We first briefly discuss a codec system which encodes an integer number w into a self-inverting permutation π, and then we present a transformation of a self-inverting permutation into 2D and 1D representations.

2.1 Self-inverting Permutations

In a formal (i.e., mathematical) way, a permutation of a set of objects S is defined as a bijection from S to itself, that is, a map $S \to S$ for which every element of S occurs exactly once as image value.

Permutations may be represented in many ways, where the most straightforward is simply a rearrangement of the elements of the set $N_n = \{1,2,\ldots,n\}$; for example, the permutation $\pi = (4,7,6,1,5,3,2)$ is a rearrangement of the elements of the set N_7 (Sedgewick and Flajolet, 1996; Golumbic, 1980).

Definition 2.1.1. Let $\pi = (\pi_1,\pi_2,\ldots,\pi_n)$ be a permutation over the set N_n, where $n > 1$. The inverse of the permutation π is the permutation $q = (q_1,q_2,\ldots,q_n)$ with $q_{\pi_i} = \pi_{q_i} = i$. A *self-inverting permutation* (or, for short, SiP) is a permutation that is its own inverse, that is $\pi_{\pi_i} = i$.

There are several systems that correspond integer numbers into permutations (Sedgewick and Flajolet, 1996). Recently, we have proposed algorithms for such a system which efficiently encode an integer w into a self-inverting permutation π and efficiently decode it; our algorithms run in $O(n)$ time, where n is the length of the binary representation of w.

2.2 2D and 1D Representations

In the 2D representation, the elements of the permutation $\pi = (\pi_1,\pi_2,\ldots,\pi_n)$ are mapped in specific cells of an $n \times n$ matrix A as follows:

$$\text{number } \pi_i \longrightarrow \text{entry } A(\pi_{\pi_i}^{-1},\pi_i)$$

or, equivalently, the cell at row i and column π_i is labeled by the number π_i, for each $i = 1,2,\ldots,n$.

Figure 1(a) shows the 2D representation of the self-inverting permutation $\pi = (4,7,6,1,5,3,2)$.

Based on the previously defined 2D representation of a permutation π, we next propose a two-dimensional marked representation (2DM representation) of π which is an efficient tool for watermarking images. In our 2DM representation, a permutation π over the set N_n is represented by an $n \times n$ matrix A^* as follows:

- the cell at row i and column π_i is marked by a specific symbol, for each $i = 1,2,\ldots,n$;

where, in our implementation, the used symbol is the asterisk, i.e., the character "*". Figure 1(b) shows the 2DM representation of the permutation $\pi = (4,7,6,1,5,3,2)$.

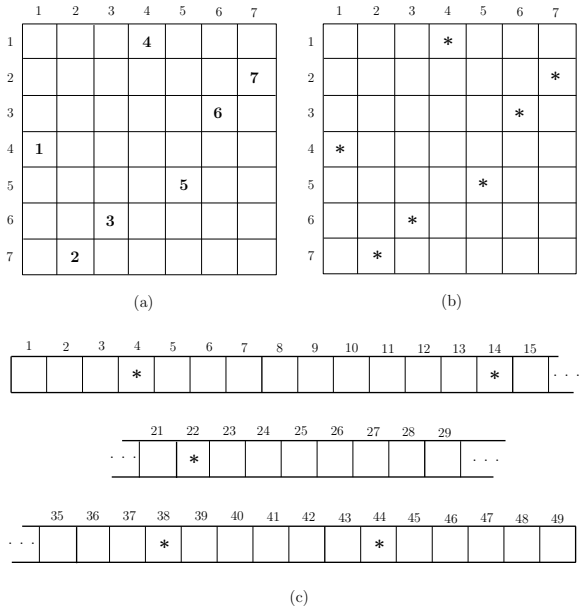


Figure 1: The 2D, 2DM and 1DM representations of the self-inverting permutation $\pi = (4,7,6,1,5,3,2)$.

In our 1D representation, the elements of the permutation π are mapped in specific cells of an array B of size n^2 as follows:

$$\text{number } \pi_i \longrightarrow \text{entry } B((\pi_{\pi_i}^{-1} - 1)n + \pi_i)$$

or, equivalently, the cell at the position $(i-1)n+\pi_i$ is labeled by the number π_i, for each $i = 1,2,\ldots,n$.

We next describe the 1DM representation acquired in a similar manner as the 2DM representation. In our 1DM representation, a permutation π over the set N_n is represented by an n^2 array B^* as follows:

- the cell at the position $(i-1)n+\pi_i$ is marked by a specific symbol, for each $i = 1,2,\ldots,n$;

where, in our implementation, the used symbol is again the asterisk character "*". Figure 1(c) shows the 1DM representation of the same permutation $\pi = (4,7,6,1,5,3,2)$.

Hereafter, we shall denote by π^* a self-inverting permutation and by n^* the number of elements of π^*.

3 PREVIOUS RESULTS ON IMAGE WATERMARKING

In a recent work of ours, we have proposed an image watermarking technique that embeds watermarks into digital images by interfering in the frequency domain of images. Since our audio watermarking technique, that is going to be later described, is mainly based on the idea of image watermarking, we next briefly describe the main steps of our image watermarking technique and state points regarding some of its main characteristics.

The embedding image watermarking algorithm first computes the 2DM representation of the permutation π^*, that is, the $n^* \times n^*$ array A^* (see, Subsection 2.2). Next, it takes the input image I, covers it with an $n^* \times n^*$ imaginary grid C, resulting in $n^* \times n^*$ grid-cells C_{ij}, and takes the Discrete Fourier Transform (DFT) F_{ij} of each C_{ij}. The algorithm goes to each grid-cell C_{ij}, takes the magnitude M_{ij}, and places on it two imaginary ellipsoidal annuli denoted as "Red" and "Blue". It then computes the average of the magnitude values grouped by the "Red" and the "Blue" annuli, say, $AvgR_{ij}$ and $AvgB_{ij}$, respectively, and after that, for each M_{ij} computes the value $D_{ij} = |AvgB_{ij} - AvgR_{ij}|$ if $AvgB_{ij} < AvgR_{ij}$, otherwise $D_{ij} = 0$. Subsequently, the algorithm computes for each row i the maximum value $MaxD_i$. Once again the embedding algorithm goes to each grid cell C_{ij} and if $A_{ij} =$ "$*$" it increases the values of M_{ij} grouped by the "Red" annulus by $AvgB_{ij} - AvgR_{ij} + MaxD_i + c_{opt}$. Finally, it reconstructs each DFT cell F_{ij} using the modified M_{ij} with the trigonometric formula and with the inverse DFT it reconstructs the grid cells C_{ij}. The extracting algorithm works in a similar manner.

Regarding the main characteristics of this technique, we should first mention that it is efficient. As the experimental results showed, watermarks are imperceptible leading also to high fidelity. Moreover, watermarks are robust to distortions as we got positive results testing the watermarked images against JPEG compression and other attacks.

4 THE AUDIO WATERMARKING TECHNIQUE

In this section we present an algorithm for encoding a self-inverting permutation π^* into an audio signal S by marking specific time segments of S in the frequency domain resulting thus the watermarked audio signal S_w. We also present a decoding algorithm which extracts the embedded permutation π^* from S_w by locating the positions of the marks in S_w.

4.1 Embed Watermark into Audio

The embedding algorithm of our proposed technique encodes a self-inverting permutation (SiP) π^* into a digital audio signal S. Recall that, the permutation π^* is obtained over the set N_{n^*}, where $n^* = 2n+1$ and n is the length of the binary representation of an integer

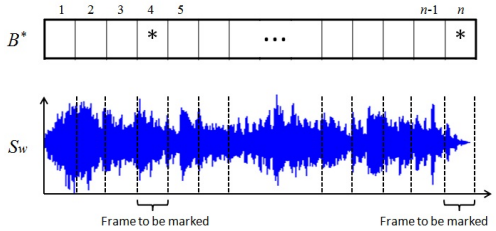


Figure 2: Segmentation of the S's signal into specific frames according to 1DM representation of the permutation π^*.

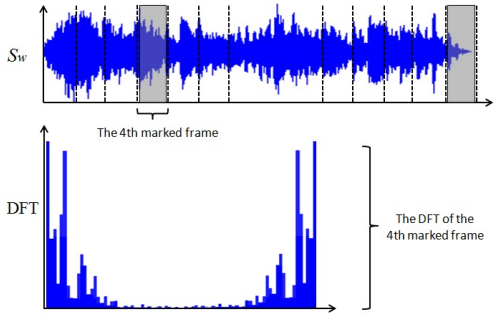


Figure 3: The DFT representation of a marked frame.

w which actually is the audio's watermark (author's technique).

The Main Idea of embedding. The watermark w, or equivalently the corresponding self-inverting permutation π^*, is imperceptibly inserted in the frequency domain of specific frames on the audio track signals S; see, Figure 2. More precisely, we mark certain frames getting the DFT and do alterations at the magnitude values of high frequencies for each audio frame to be marked; see, Figure 3. This is achieved by choosing two groups of magnitude values specified with two segments of the magnitude vector namely "Red" and "Blue" and the alterations are actually on their difference; see, Figure 4. In our implementation we use fixed segments' widths and distances from the center of symmetry of the DFT's magnitude vector. The added value is specified by the maximum value in the defined area.

The Embedding Algorithm. Our embedding algorithm takes as input a SiP π^* and an audio signal S and returns the watermarked audio signal S_w; it performs the following main processes:

i. construct the 1DM representation of the watermark number w;

ii. transform the input audio signal S and acquire the frequency representation of it;

iii. modify signals' frequency representation according to the 1DM representation of the signal S;

iv. returns the watermarked audio signal S_w;

We describe below in detail the embedding algorithm in steps.

Algorithm: Embed_SiP-to-Audio.
Input: the watermark $\pi^* \equiv w$ and the original audio signal S;
Output: the watermarked audio signal S_w;

Step 1: Compute first the 1DM representation of the permutation π^*, i.e., construct the array B^* of size $n = n^* \times n^*$; recall that the entry $B^*((i-1)n^* + \pi_i^*)$ contains the symbol "*", $1 \leq i \leq n^*$.

Step 2: Segment the audio signal S into n non-overlapping frames f_i of size $f_i[a,b] = \lfloor \frac{N-1}{n} \rfloor$, $1 \leq i \leq n$, where N is the length of the audio signal.

Step 3: For each frame f_i, compute the Discrete Fourier Transform (DFT) using the Fast Fourier Transform (FFT) algorithm, resulting in n DFT frames F_i of size $F_i[a,b] = \lfloor \frac{N-1}{n} \rfloor$, $1 \leq i \leq n$, that is, $F_i = \texttt{FFT}(f_i)$.

Step 4: For each DFT frame F_i, compute its magnitude M_i and phase P_i vectors (or, arrays) which are both of size $M_i[a,b] = P_i[a,b] = \lfloor \frac{N-1}{n} \rfloor$, $1 \leq i \leq n$.

Step 5: Then, the algorithm takes each of the n magnitude vectors M_i and determines two segments in M_i, $1 \leq i \leq n$, denoted as "Red" and "Blue" (see, Figure 4). In our implementation,

- each "Red" segment $[x_r, y_r]$ has length ℓ_r (even), where $x_r = \lfloor \frac{N-1}{2n} \rfloor - \frac{\ell_r}{2}$ and $y_r = \lfloor \frac{N-1}{2n} \rfloor + \frac{\ell_r}{2}$;
- each "Blue" segment $[x_b, y_b]$ has length ℓ_b (even), where $x_b = x_r - \frac{\ell_b}{2}$ and $y_b = y_r + \frac{\ell_b}{2}$

The "Red" and the "Blue" segments determine two groups of magnitude values on M_i; the Red_Values and the Blue_Values (see, Figure 4).

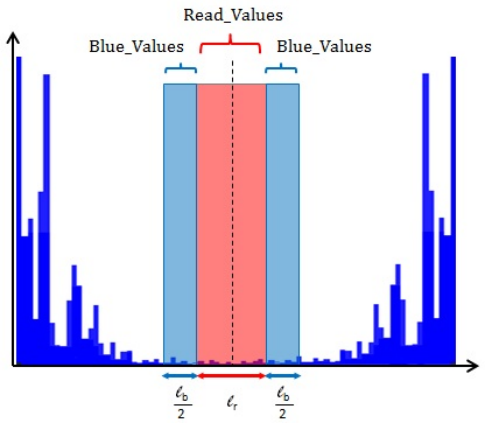


Figure 4: The "Red" and "Blue" segments on DFT.

Step 6: For each magnitude vector M_i, $1 \leq i \leq n$, compute the average value $AvgR_i$ of the Red_Values and the average value $AvgB_i$ of the Blue_Values of M_i.

Step 7: For each magnitude vector M_i, $1 \leq i \leq n$, compute first the variable D_i as follows:

- $D_i = |AvgB_i - AvgR_i|$, if $AvgB_i \geq AvgR_i$
- $D_i = 0$, otherwise.

Step 8: Partition the n values D_1, D_2, ..., D_n into n^* sets $E_1, E_2, \ldots, E_{n^*}$, each of size n^* (recall that $n = n^* \times n^*$); let $\{D_{i1}, D_{i2}, \ldots, D_{in^*}\}$ be the elements of the i-th set E_i, $1 \leq i \leq n^*$. Then, compute the values

- $MaxD_1, MaxD_2, \ldots, MaxD_{n^*}$

where $MaxD_i$ is the maximum value of the i-th set $E_i = \{D_{i1}, D_{i2}, \ldots, D_{in^*}\}$, $1 \leq i \leq n^*$.

Step 9: For each marked cell $B^*(i)$ of the 1DM representation matrix B^* of the permutation π^* (i.e., the call which contains the symbol "*"), mark the corresponding frame F_i, $1 \leq i \leq n$; the marking is performed by increasing all the Red_Values in M_i by the value

$$AvgB_i - AvgR_i + MaxD_k + c, \tag{1}$$

where $k = \lceil \frac{i}{n^*} \rceil$ and $c = c_{opt}$. The additive value of c_{opt} is a predefined value which enables successful extracting.

Step 10: Reconstruct the DFT of the corresponding modified magnitude vector M_i, using the trigonometric form formula (Gonzalez and Woods, 2007), and then perform the Inverse Fast Fourier Transform (IFFT) for each frame F_i, $1 \leq i \leq n$, in order to obtain the audio signal S_w.

Step 11: Return the watermarked audio signal S_w.

Note that concerning the placement of the "Red" and "Blue" segments, their position can vary according to the frequency band in which we want to mark a frame. At the above illustration we mark it in the high frequencies but there can be a different approach. Specifically, we can mark instead lower frequencies and that is performed by moving the segments from the center to the right and left edges of the magnitude array of the Discrete Fourier Transform (DFT) representation.

4.2 Extract Watermark from Audio

In this section we describe the decoding algorithm of our proposed technique. The algorithm extracts the SiP π^* from a watermarked digital audio signal S_w, which can be later represented as an integer w.

The Main Idea of Extracting. The main idea behind the extracting algorithm is that the self-inverting permutation π^* is obtained from the frequency domain of specific frames of the watermarked audio signal S_w. More precisely, using the same two "Red" and "Blue" segments, we detect certain areas of the watermarked audio signal S_w so that the difference between the average values of the "Red" segment have the maximum positive difference over the average values of the "Blue" segments. In this way we can detect marked frames that enable us to obtain the 1DM representation of the permutation π^*.

The Extracting Algorithm. We next describe the extracting algorithm which consists of the following steps.

Algorithm: Extract_SiP-from-Audio.
Input: the watermarked audio S_w marked with π^*;
Output: the watermark $\pi^* = w$;

Step 1: Take the input watermarked audio S_w and compute its size N. Then, segment S_w into n non-overlapping frames f_i of size $f_i[a,b] = \lfloor \frac{N-1}{n} \rfloor$, $1 \leq i \leq n$.

Step 2: Then, using the Fast Fourier Transform (FFT), get the Discrete Fourier Transform (DFT) for each frame f_i, resulting in n DFT frames F_i, $1 \leq i \leq n$.

Step 3: For each DFT frame F_i, compute its magnitude M_i and phase P_i vectors, which are both of size $M_i[a,b]$=$P_i[a,b]$=$\lfloor \frac{N-1}{n} \rfloor$, $1 \leq i \leq n$.

Step 4: For each magnitude vector M_i, compute the average values $AvgR_i$ and $AvgB_i$ of the Red_Values and Blue_Values of M_i, respectively, as described in the embedding algorithm.

Step 5: Partition the n vectors M_i, $1 \leq i \leq n$, into n^* sets $L_1, L_2, \ldots, L_{n^*}$, each of size n^*; let $\{M_{i1}, M_{i2}, \ldots, M_{in^*}\}$ be the elements of the i-th set L_i and let $AvgR_{ij}$ and $AvgB_{ij}$ be the average values of the Red_Values and Blue_Values, respectively, of the vector M_{ij}, $1 \leq i, j \leq n^*$.

Step 6: For each set $L_i = \{M_{i1}, M_{i2}, \ldots, M_{in^*}\}$ find the k_{th} vector M_{ij} such that $AvgB_{ik} - AvgR_{ik}$ is minimum and set $\pi^*_i = k$, $1 \leq k \leq n^*$.

Step 7: Return the self-inverting permutation π^*.

Having presented the embedding and extracting algorithms, we next briefly comment on the purpose of the additive value $c = c_{opt}$ (see, Step 9 of the embedding algorithm). Similar to image watermarking, we add at the corresponding embedding marking step the additive value c_{opt} which by getting greater increases the

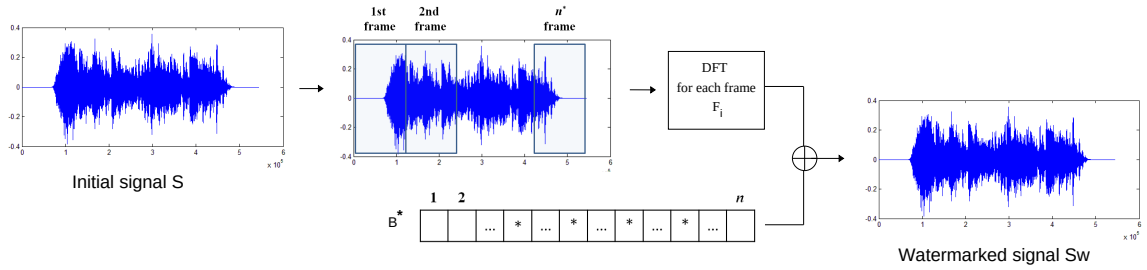


Figure 5: The encoding process of audio signal watermarking.

robustness of the marks; in our audio watermarking case, we just used a very small value for it.

5 EXPERIMENTAL RESULTS

This section summarizes the experimental results of the proposed audio watermarking codec algorithms; we implemented our algorithms and carried out tests using the general-purpose mathematical software package Matlab (Version 7.7.0) (Ingle and Proakis, 2010).

Testing of our embedding and extracting algorithms has been made by the use of various 16-bit digital audio tracks in *wav* format with 44.1 KHz sampling frequency. Concerning the audio samples used, they where relatively short abstracts with different characteristics. For instance there were tracks containing speech which have many silent segments as well as music track samples and tracks with extreme features such as low and high frequency sounds. Many of the audio tracks that we used for testing were acquired from a web audio repository called wavsource and enriched by some other audio tracks from various sources.

It is well known in the field of watermarking that there are three main characteristics to take into account describing and evaluating a digital watermarking system: *Fidelity*, *Robustness*, and *Capacity* (Cox et al., 2008).

Concerning our watermarking system, it seems to be of high fidelity as watermarked tracks were not distinguished over the original ones and the results using the PSNR metric were interestingly positive.

Concerting the marking procedure of our implementation, we set both lengths ℓ_r and ℓ_b of the "Red" and "Blue" segments respectively, equal to 20% of half the length of magnitude vector as it is mirrored (see, Section 4.1). Recall that, the value 20% is a relatively small percentage which allows us to modify the audio track segments in a satisfactory level in order to detect the watermark and successfully extract it without affecting audio tracks' initial quality. Moreover, we choose to alter higher frequencies and thus the two segments are at the center of the magnitude vector. This is because high frequencies are less perceptible according to the human auditory system. What is more, at high frequencies audio tracks contain less information which means that information is less likely to be lost due to post alterations.

Fidelity. In order to evaluate the watermarked audio track quality obtained from our proposed watermarking method we used the Peak Signal to Noise Ratio (PSNR) metric. Our aim was to prove that the watermarked audio track is closely related to the original track proving the high fidelity attribute of our system. This is something vital as watermarking should not introduce audible distortions in the original audio track, as that would certainly reduce its commercial value.

Giving a short introduction to the PSNR metric, we should mention that it is defined as the ratio between the reference (or, original) signal and the distorted (or, watermarked) signal of an audio track and it is given in decibels (dB). It is well known that PSNR is most commonly used as a measure of quality of reconstruction of lossy compression codecs (e.g., for image or audio compression methods). The higher the PSNR value the closer the distorted signal is to the original or the better the watermark conceals. We mentions that PSNR is a popular metric due to its simplicity.

For an initial audio signal S of size N and its watermarked equivalent signal S_w, PSNR is defined by the formula:

$$\mathrm{PSNR}(S,S_w) = 10\log_{10}\frac{N_{max}^2}{MSE}, \qquad (2)$$

Table 1: The hamming distance of the watermark w and the extracted watermark w^* after common signal attacks.

Filename	Gaussian Noise	Cropping	Resampling	Requantization	MP3 Compression
bach.wav	0	1	0	0	3.2
clarinet.wav	0	1	0	0	2.4
castanets.wav	0	1	0	0	5.2
elvis_riverside.wav	0	1	0	0	2.8
family_man.wav	0	1	0	0	0.2
high10sec.wav	0	1	0	0	3.0
low10sec.wav	0	1	0	0	2.0

where N_{max} is the maximum signal value that exists in the original audio track and MSE is the Mean Square Error which is given by the following formula:

$$\mathrm{MSE}(S,S_w) = \frac{1}{N}\sum_{i=0}^{N-1}(S(i)-S_w(i))^2. \qquad (3)$$

Comparing the original audio tracks with the watermarked ones, we immediately get to notice that they depict excellent fidelity according to the PSNR values that we have obtained. In every case PSNR is over 50 dB which proves that fact that there is a striking similarity between the original and the watermarked signal of an audio track.

Table 2: The PSNR values of the watermarked audio signals.

Filename	PSNR
bach.wav	67.2
clarinet.wav	67.9
castanets.wav	68.2
elvis_riverside.wav	75.3
family_man.wav	73.8
high10sec.wav	58.8
low10sec.wav	64.5

In Table 2 you can see the performance of our method as we demonstrate the PSNR values of some audio tracks that we used in this work. Each of them was sampled at 41.1 KHz and the duration was of about 10 sec. Additionally, each one has much different characteristics. More specifically, the audio tracks

- bach.wav, clarinet.wav and castanets.wav

where a concert, a clarinet and castanets solo respectively. The audio track

- elvis_riverside.wav

combines human voice with music, while the

- family_man.wav

contains only speech which means that it also has periods of silence. Lastly the audio tracks

- high10sec.wav and low10sec.wav

are some extreme cases of high and low frequency sounds.

Robustness. The watermarked signals were subjected to distortions or common signal attacks in order to evaluate the robustness of our audio watermarking algorithms. We tested the performance of each audio track under white noise addition, cropping, resampling, requantization and MP3 compression. Below we describe in more details each one of the five different attacks that we applied in our experiments.

(a) Gaussian Noise. A white gaussian noise of SNR 20 dB was added to the original audio signal.

(b) Cropping. A 10% of the beginning of the watermarked audio signal was cropped and subsequently replaced by zeros.

(c) Resampling. The watermarked signal, originally sampled at 44.1 KHz, is resampled at 22.05 KHz, and then restored back by sampling again at 44.1 KHz

(d) Requantization. The 24-bit watermarked audio signal is re-quantized down to 16 bits/sample and then back to 24 bits/sample.

(e) MP3 compression. The watermarked audio signal is compressed using a bit rate of 128 Kb/s and then decompressed back to the WAV format.

Since the watermark that we embed in our audio signal is a permutation, i.e. a vector over the set N_n ($n > 1$), we test after each attack the similarity of the extracted watermark with the original one using the Hamming distance (Hamming, 1950).

The Hamming distance $d(x,y)$ between two vectors x and y is the number of coefficients in which they differ (Hamming, 1950). The Hamming distance equals to zero, i.e., $d(x,y) = 0$, if x and y agree in all coordinates; it happens if and only if $x = y$. In our case the Hamming distance is computed between the watermark $w = \pi^*$ that we embedded into the audio track and the watermark π^*_{ext} that we extract from the audio. If $d(\pi^*, \pi^*_{ext}) = 0$ the watermark $w = \pi^*$ successfully

extracts from the attacked audio signal. Additionally, it is worth noting that if $d(\pi^*, \pi^*_{ext})$ is relatively small, then the watermark π^* can be reconstructed with high probability by exploiting the self-inverting properties of the permutation π^*.

In Table 1 we demonstrate similarity results between the watermark that we embedded into the audio track and the watermark that we extracted after various signal processing attacks. As the experimental results show, our audio watermarking algorithm is robust against additive gaussian noise of SNR 20 dB, cropping, resampling and requantization. Evaluating our method's robustness over lossy compression we tested it using the MP3 encoding format with a bit rate of 128 Kb/s. In order to optimize the results as high frequency information is mostly lost using MP3 we made the appropriate adjustments concerning the width of the segments to be marked as well as the additive value $c = c_{opt}$ (see, Algorithm Embed_SiP-to-Audio). For most cases the results were positive as despite not being able in every case to successfully extract all the elements of the watermark, using the properties of self-inverting permutations recovery of the initial watermark can be successfully operated.

Closing the robustness evaluation of our method, we should point out that a drawback of our method is actually when we want to watermark an audio track with extreme high frequencies; it is something that could be encountered on future work.

Capacity. The capacity of our audio watermark method has been computed by measuring the percentage of the watermarked parts of an audio track over the length of the entire audio track. Our method partitions the audio track into $n^* \times n^*$ frames, where n^* is the length of the permutation π^*, and marks only one frame of a set of n^* frames; recall that our embedding method groups the frames into n^* sets each containing n^* frames (see, Algorithm Embed_SiP-to-Audio). That means, a total n^* over $n^* \times n^*$ frames are marked, so the ratio of the watermarked part over the entire length of the audio track is $\frac{n^*}{(n^* \times n^*)}$. Thus, our audio watermarking method has $\frac{1}{n^*}$ capacity.

6 CONCLUDING REMARKS

In this paper we presented an audio watermarking technique which efficiently and invisibly embeds information, i.e., watermarks, into an audio digital signal. Our technique is based on the same main idea of a recently proposed image watermarking technique expanding thus the digital objects that can be efficiently watermarked through the use of self-inverting permutations.

We experimentally tested our embedding and extracting algorithms on WAV audio signals. Our testing procedure includes the phases of embedding a numerical watermark $w = \pi^*$ into several audio signals S, storing the watermarked audio S_w in WAV format, and extracting the watermark $w = \pi^*$ from the audio S_w. We obtained positive results as the watermarks were invisible, they didn't affect the audio's quality and they were extractable.

The performance evaluation of our audio watermarking technique on several other attacks remains a problem for further investigation.

REFERENCES

Alsalami, M. A. and Al-Akaidi, M. M. (2003). Digital audio watermarking: survey. *De Montfort University*, pages 1–14.

Bender, W., Gruhl, D., and Morimoto, N. (1996). Techniques for data hiding. In *Proc. IBM systems journal*, volume 35(3.4), pages 313–336.

Collberg, C. and Nagra, J. (2010). *Surreptitious Software*. Addison-Wesley.

Cox, I. J., Miller, M. L., Bloom, J. A., Fridrich, J., and Kalker, T. (2008). *Digital Watermarking and Steganography*. Morgan Kaufmann, 2nd edition.

Golumbic, M. (1980). *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, Inc., New York.

Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing*. Prentice-Hall, 3rd edition.

Grover, D. (1997). *The Protection of Computer Software - Its Technology and Applications*. Cambridge University Press, New York.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.

Hartung, F. and Kutter, M. (1999). Multimedia watermarking techniques. In *Proceedings of the IEEE*, volume 87(70), pages 1079–1107.

Ingle, V. K. and Proakis, J. G. (2010). *Digital Signal Processing using Matlab*. Cengage Learning, 3rd edition.

Sedgewick, R. and Flajolet, P. (1996). *An Introduction to the Analysis of Algorithms*. Addison-Wesley.

Sharma, S., Rajpurohit, J., and Dhankar, S. (2012). Survey on different level of audio watermarking techniques. *Int'l Journal of Comput. Applications*, 49(10):41 – 48.

Implicit User Profiling in News Recommender Systems

Jon Atle Gulla[1], Arne Dag Fidjestøl[1], Xiaomeng Su[2] and Humberto Castejon[2]
[1]*Dep. of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway*
[2]*Telenor Group, Trondheim, Norway*
{jag, adf}@idi.ntnu.no, {xioameng.su, humberto.castejon}@telenor.com

Keywords: Recommender Systems, Personalization, User Profiling, Mobile News, Big Data, Information Retrieval.

Abstract: User profiling is an important part of content-based and hybrid recommender systems. These profiles model users' interests and preferences and are used to assess an item's relevance to a particular user. In the news domain it is difficult to extract explicit signals from the users about their interests, and user profiling depends on in-depth analyses of users' reading habits. This is a challenging task, as news articles have short life spans, are unstructured, and make use of unclear and rapidly changing terminologies. This paper discusses an approach for constructing detailed user profiles on the basis of detailed observations of users' interaction with a mobile news app. The profiles address both news categories and news entities, distinguish between long-term interests and running context, and are currently used in a real iOS mobile news recommender system that recommends news from 89 Norwegian newspapers.

1 INTRODUCTION

With the growing popularity of smartphones and tablets an abundance of news apps have been introduced over the last few years as alternatives to online web news sites. News aggregator apps like News360, Flipboard, Pulse and Feedly are not linked to one particular media house, but make use of publically available news stories as soon as they are published on the Internet. These mobile applications allow users to read neatly presented news stories from numerous media houses using interfaces that require only very limited interaction with the system. The features of these apps vary, though information filtering and user friendliness are key features of many of these applications (Haugen, 2013).

Advanced recommendation technologies are still rare in commercial news apps, but some of them now include simplified recommendation features, and several research prototypes experiment with new and promising recommendation approaches that analyze both the content of news articles and the users' social networks. This is not very surprising, as mobile news may benefit from these recommendation technologies for a number of reasons: (i) the constant flow of news easily leads to information overload problems for the reader, (ii) the small screen prevents the apps from showing several stories or proper news overviews, and (iii) the lack of an efficient keyboard hampers the reader's interaction with the news apps.

However, an efficient recommender system needs some understanding of the users' preferences and interests. Some news apps do not keep any particular information about the user and there is no user-tailoring of news or recommendations offered. Other apps allow the user to select a category or post a query. The category or query associated with the user is then the simplest form of user profile found in mobile news apps. For full-fledged recommender systems, though, a more comprehensive representation of user interests and preferences is needed to provide personalized news services. If news stories read in the past is a good indication of the user's current preferences, we need to analyze these stories and find ways of capturing their content in her user profile. If stories read by friends or similar people are assumed to be relevant, we need ways of relating the user's profile to other users' profiles and find exactly those other users that have read news stories that are most likely to be of interest to her. There are many aspects of user profiles, and the choice of recommendation technique also influences the structure, content and maintenance of these profiles.

The NTNU SmartMedia project is experimenting with advanced recommendation techniques in a

news app for Norwegian newspapers. Users are in this app associated with devices, and the mobile news recommender engine ranks news using a combination of freshness, geographical proximity and match with users' preferences. Central in the work is the definition of user profiles that reflect and interpret user behavior and support both content-based and collaborative filtering.

The purpose of this study is to investigate to what extent a minimalistic swipe-based mobile news interface can be used to infer user preferences without any explicit signals from the users. The paper itself is structured as follows: After the introduction in Section 1, we discuss the particular challenges and needs in news recommender systems in Section 2. In Section 3 we briefly present earlier work on user profiling in recommender systems. Section 4 introduces the SmartMedia news recommendation project and its overall method for extracting user profiles from end user's interaction with the mobile app. After discussing the way this profiling technique can be configured to serve different purposes in Section 5, we conclude the paper in Section 6.

2 NEWS RECOMMENDATION

A news recommender system filters incoming news and presents to each individual user a ranked list of news that it deems most relevant to the particular user at hand. Since these are articles that the users have not already read and evaluated, we need to find a way of guessing users' interests based on their previously observed behavior.

In formal terms, the task in news recommendation is to estimate and rank the evaluations of news articles unknown to a user (Borges & Lorena, 2010; Jannach et. al, 2010). Assume a set of users U and a set of news articles A, the recommender system needs a utility function s that defines the evaluation v of an article a for a user u:

$$s: U \times A \rightarrow V,$$

in which V is a completely ordered set formed by non-negative values within an internal, e.g. 0 to 1 or 0 to 100. The system will recommend an article a' that maximizes the utility function for the user:

$$a' = \arg\max_{a \in A} s(u, a)$$

For those articles that have not been evaluated by the user, the system needs to estimate their evaluations from relatedness with other articles that have in fact been read. Techniques like decision trees, Bayesian classifiers, support vector machines, singular value decomposition, clustering and various similarity scores have been used as part of this estimation process.

In content-based filtering, which is the primary concern of this paper, the estimation is all based on the content of previously read news articles. The assumption is that users read articles on topics they find interesting, and the users' interests do not change substantially from one day to another. If a particular user preferred to read about politics one day, chances are good that she will also be interested in political news the next day. Moreover, the degree of interest in a topic may be reflected in the frequency with which the user visits news stories of this topic or category.

2.1 News Challenges

Recommender systems for products and services normally deal with items that are rather stable both in terms of number and descriptions. The recommendation task is to compare a fixed number of items that all have structured and well understood properties. There is no relevant temporal dimension, and there is no problem separating one product or service from another.

The news domain is intrinsically more dynamic and unpredictable to deal with. There are new stories coming up all the time. Some of these uncover new events, while others just report on the progress of an already published event. There may be several conflicting stories of the same event, but there may also be several events discussed in the same story.

For news recommendation there are three particular challenges that need to be addressed:

First, the news domain is characterized by fluctuating and unclear vocabularies and ever changing news topics. Rather than ranking a fixed number of products, the system need to rank a dynamically growing number of events that may have very little in common with what the user read yesterday. The vocabularies change as new stories emerge, making it difficult to detect how stories are related and thereby estimate their relevance to the user.

Second, the short life span of news stories renders most news stories irrelevant, even though they seem consistent with the user's interests. For a developing story, it makes sense to recommend the latest article, unless there are substantial differences in quality and depth that make an older article more

informative. On the other hand, if collaborative filtering is used, it may be difficult to pick up the latest article before anyone else in your community has read and rated it.

Third, most media houses consider serendipity to be important to hold on to its readers. Serendipity refers to the introduction of news that may not have been selected or preferred by the user, but may still be interesting because they are surprising, alarming, important, have a particular journalistic quality, or simply add variation into the news stream. Serendipitous news help media houses expand users interests in their products, since they persuade users to read stories that are alien to them and that would never be read if the user only received stories consistent with their reading history.

While other recommender systems concentrate on the utility function itself, news recommender systems need also to consider content analysis techniques to extract and structure the topics pertaining to the news. They need to deal with ambiguities, uncertainties and inconsistencies that affect the way articles should be formally represented and compared to users' interest profiles.

3 RELATED WORK

A user profile is a representation of user's interests and preferences that is used to verify to what extent news stories are relevant to a particular user. The profiles are built for each individual user, are regularly updated, and describe topics, news categories and relevant features of the users.

In principle, there are two types of user profiles, *profiles based on implicit feedback* and *profiles based on explicit feedback* (Gauch et al, 2007; Lops et. al, 2011). Implicit user profiles are automatically extracted by the recommender systems themselves and may or may not be a correct representation of users' interests. In general, implicit feedback methods assign relevance scores to user actions on news articles like saving, sharing, bookmarking, etc. Explicit profiles are entered or approved by the users in question, but tend to be slightly less detailed than the implicit ones. The user-selected categories in Zite and Flipboard form very simple explicit user profiles, while the thumbs up/down approach in Zite and News360 provide explicit signals for what topics should be included in the profiles. However, adopting numeric or symbolic scales increases the users' cognitive load and may not be adequate for capturing emotions or attitudes towards the news. For mobile news recommendation it seems difficult to require that the user enter and regularly update extensive representations of her interests, though, and more advanced techniques for profile construction on the basis of implicit feedback are needed.

A system combining explicit feedback and automatic learning is described in Singh et. al (2006). After building an initial interest category hierarchy on the basis of explicit feedback on a number of articles, the system analyzes user feedback from ongoing news sessions and automatically adds new leaf categories or update existing ones in the interest hierarchy. A similar approach is taken by Kim & Chan (2003), though they depend less on explicit user feedback.

Billsus and Pazzani (2000) have developed an approach in DailyLearner for interpreting implicit user feedback on news articles. A user click on the headline of an article is taken as a signal of interest, and an initial score of 0.8 is set. If the user is requesting more pages of the story, the score will be gradually increased until it reaches a maximum of 1.0 when all pages have been consulted. Similarly, a skipped article is assumed to be of no interest and is given a negative score that is subtracted from the system's prediction score for the article. All these scores are combined into a user profile that lists weighted informative words typical for the user's interests and preferences.

Liu, Dolan and Pedersen (2010) build user profile vectors that express users' interests in specific news categories over time. For each user they record the distribution of clicks and associate these click rates with categories on a monthly basis. This allows them to assess for every user the proportion of time spent on reading news from each category as well as to reflect on the development of her interests from one month to another.

In Cantador et. al (2008) they use semantic expansion in combination with a standard user preference algorithm for content-based recommendation. They observe that automatically learned profiles tend to be dominated by the main characteristics of user's preferences, preventing the recommendation engine from recommending news that are related albeit not directly addressed by the profile. Their solution is to use an ontology to include additional weighted concepts in the profile that are related to the original learned concepts or terms. Then they combine this expanded profile with a running context, which is a weighted set of concepts from the user's latest interactions with the system – to produce a contextualized version of the user's preferences that filters out topics that are out

of focus.

Implicit signals of user interests require that the user interface is designed in such a way that different levels of interests in a news article are reflected in the user's interaction with the system. This is a challenge in itself, as most news apps try to minimize the required interaction between system and reader. More interaction is good for analysing user interests, though it may lead to a less attractive user experience.

4 THE SMARTMEDIA PROJECT

The SmartMedia project at NTNU was initiated in 2011 as part of the university's collaboration with the Norwegian media industry and major Telecom companies. The project targets recommendation technologies for the mobile news domain and addresses content-based and collaborative techniques in combination with semantic and linguistic theories (in Gulla et. al (2014).

Another important issue is the use of platforms, frameworks and methods for designing efficient cross-platform mobile user interfaces. Figure 1 shows the current mobile news recommender systems for the iOS platform. Details of this implementation can be found in Tavakolifard et. al (2013). A new version of the system in HTML5/JavaScript is under development and will introduce new features for hybrid news recommendation.

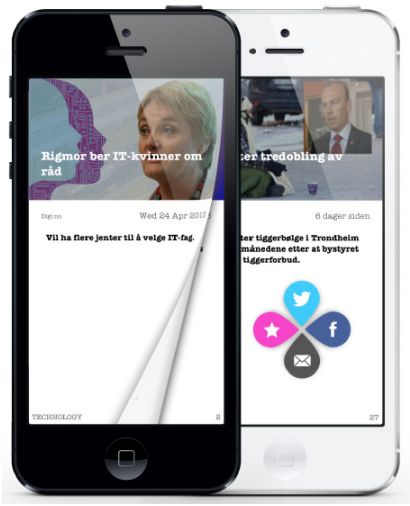


Figure 1: SmartMedia's news recommender app.

The current iOS news app indexes about 1,500 news articles per day. The articles are sourced from 89 Norwegian newspapers and are mostly in one of Norway's two official languages. During the indexing process, we extract and index the article's category (e.g. sports or lifestyle), its location (e.g. Oslo), and relevant keywords characterizing the article's content. The keywords are partly named entities (e.g. Lionel Messi) from a NER (named entity recognition) component, partly important nominals that refer to concepts discussed in the article. The underlying news index is in SolR, an open source search platform from the Apache Lucene project, and is built up as a traditional large-scale search engine index (see for example Gulla et. al (2002) and Solskinnsbakk & Gulla (2010)).

5 USER PROFILING

In SmartMedia's user profiling component a set of 10 user acts have been identified as implicit signals of user interest. These are *Opened article view, Article view time, Preview time, Clicked category, Shared twitter, Shared facebook, Shared mail, Starred article, Viewd map,* and *Viewed similar article.* We monitor the time spent reading a news story or looking at its preview, assuming that large view dwell times indicate a stronger interest in the topic.

The construction of a user profile for a particular user follows a three-step process:

1. *Extract interests from user acts*
2. *Build running context from all user acts*
3. *Combine running context and long-term interests into new user profile*

In the following, we go into detail of each of these three steps.

5.1 User Acts

Whenever the user is interacting with the news app, her actions are logged and stored with all relevant data associated with the act. This includes not only the user ID and the act itself, but also the time of the act, the location of the news reported in the article, the news category, as well as entities and important key phrases describing the content of the article.

Figure 2 shows how a particular user act is represented in the system. As seen from the timestamp, this is the act of reading a particular news story at 16:41 on 2nd June 2013. The user spent 1.4 seconds reading it, and the article itself was a news article about an Indian man that resisted an arrest in Kristiansand in Norway. The content is

revealed from the keywords (named entities and important nominal phrases) that are listed under the 'tags' label.

As different types of user acts may signal different levels of user interest, each type of act *t* is associated with a constant α_t . For timed user acts, like reading a preview or the full version of a news story, we only want to consider this reading if it is a significant act, i.e, she spends more time than usual on the story.


```
{
"_id" : "241B50BE-DFF5-4AAB-A12D-98D4A4606028" ,
"articleId" : "318218311" ,
"userId" : "bf4d2b7adec01da0ddc8c3317e88bcdc6" ,
"eventType" : "TIME_SPENT_ARTICLE_VIEW" ,
"timestamp" : { "$date" : "2013-06-02T16:41:15.511Z"} ,
"geoLocation" :
                { "name" : "" ,
                "type" : " " ,
                "longitude" : 8.00354 ,
                "latitude" : 58.138821} ,
"properties" : { "duration" : "1.427272"} ,
"tags" :
                [ "agder politidistrikt" ,
                "havnet" ,
                "satt" ,
                "rebelsk mannen" ,
                "operasjonsleder" ,
                "kristiansund slo" ,
                "politiet" ,
                "kristiansand skallet" ,
                "Maharashtra" ,
                "India" ,
                "Egersund" ,
                "Rogaland" ,
                "Norway" ] ,
"categories" : [ "NEWS]}
```


Figure 2: User act logging.

At the content level, we now assume that the interests of the user are reflected by the important keywords found in the article she has read, modified by the importance of the user act. We model this as a *user act interest vector* $\overrightarrow{E_t} = (e_1,..., e_k)$ for act i, where e_j is defined as

$$e_j = \begin{cases} 0 & \text{if act } i \text{ is insignificant timed act} \\ \alpha_t * f_j & \text{otherwise} \end{cases}$$

α_t denotes the importance of user acts of type t, while f_j is the frequency of keyword j. In the current implementation α_t is set to 1 for all types *t*, though experiments show that some of these acts are more significant than others.

A similar analysis is used for categories, giving us also a user act category vector.

5.2 Running Context

A running context for a user is built from all user acts of that user that have happened after the last time a full user profile was generated. It describes the overall topics of what the user has been clicking on or reading lately without reflecting what she might have been interested in at earlier occasions. The running context gives us the user's current news focus.

Formally, the *running content context* $\overrightarrow{V_R}$ is defined as follows:

$$\overrightarrow{V_R} = \sum_{i \in T} E_i$$

where i ∈T are the user acts in the time span T after the last time her profile was generated.

A similar running category context is also defined.

Implementationally, the mobile news app records every gesture from the user and maintains an updated running context at all times. Whenever the user decides to reset his context or consult her stored profile, the running context is compared with her old user profile on the server side. When the user ends her session, the running context is incorporated into her long-term user profile, so that she will start with an updated user profile and an empty running context next time.

5.3 User Profile

Our user profiles combine long-term interests with users' focus on current news stream. The idea is that the user may be interested in following news events that are unfolding, even though they are not necessary perfectly in line with her general interests and preferences.

For each user the system stores an old user profile $\overrightarrow{V_0}$. The new user profile $\overrightarrow{V_U}$ is calculated from the old profile and the running context as follows:

$$\overrightarrow{V_U} = c\overrightarrow{V_0} + (1 - c)\overrightarrow{V_R}$$

If c is set to 1, the user profile does not take into account the running context and will never change. A value of 0 makes the old profile irrelevant, and only the running context is used to recommend news stories to the user. I practice the constant c should be carefully selected to both address the users' need for relevant news and the content providers' need to promote the latest news.

As before there is a similar calculation for constructing the user profile at category level.

It is important to note that we also assume that long-term interests fade unless they are renewed when similar topics come up in the news stream. Since *c* will always be below 1, the features of old

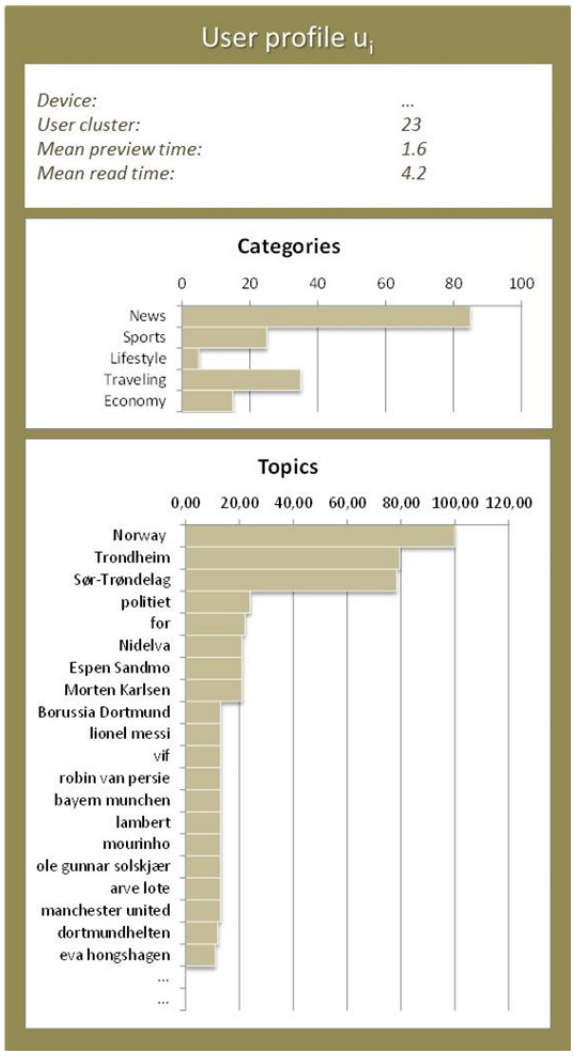


Figure 3: Generated user profile.

user profiles gradually disappear as they are updated with running contexts. The lower the *c* value is, the faster the system will forget the user's old profile.

Figure 3 shows a generated user profile for a user that has just started using the system. Both the category part and the content part of the profile are shown. The category part tells us that this user is mostly interested in news, though she has also enjoyed some stories about traveling and sports. At the content level, we notice that the user has preferred local news, as *Norway, Trondheim* and *Sør-Trøndelag* are all location names relevant to his home place. The rest of the profile exhibits some sports personalities as well as culture persons in the Norwegian media. At the top of the profile, there are two variables, *mean preview time* and *mean read time*, that are used to determine if a timed user act is significant or not. If the user spends less than 4.2 seconds reading a full news story, the system will not consider this reading act relevant to the calculation of her user profile.

6 DISCUSSION

The SmartMedia news recommender app sources news from all major newspapers in Norway. Some of these news come with meta data like category and tags, though most news stories are just unstructured of texts that need to be analyzed to locate the news, categorize the stories and extract informative keywords that describe their content. Since these keywords are critical for both the construction of profiles and subsequent recommendation process, their quality and availability are important to the recommendation process.

With the exception of finance articles, news stories in Norway are on average 220 words (tokens) long. Finance news tend to be more than 1,000 words long and written in a more factual and objective manner.

On the average our news articles contain the following named entities:

Named entity	Occurrences per story
Locations	*1.6*
Persons	*2.3*
Organizations	*2.3*
Roles	*0.8*

Since there are so few named entities in news stories, our user profiles also include important nominal phrases that constitute more concepts than references to particular entities.

The evaluation of user profiles has so far focused on the contribution of each type of user acts on the quality of the profile (Nilsen, 2013). In one experiment we used sensitivity analyses to assess the influence of each type of user acts, and to what extent they are consistent with each other. The results show that the *Opened article view* act in many cases provided a good average of the other acts taken together, though the other acts were needed to deal with particularly interesting or uninteresting news stories. Another experiment analyzed the impact on user profile size. Not surprisingly, the *Preview time* act accounted for 82% of the profile vectors' contents after some time of regular use. This may be an indication that our threshold for taking this act into account is set to low or has to reflect a more dynamic rule than just accepting any reading act longer than the average one. Quite likely the lengths of the news stories

need to be a factor when considering the reading time.

Experiments with the news recommender system shows that there are three issues that need to be carefully considered when user profiles are constructed:

- **Granularity**. In the current implementation we consider category interests and content interests equally important. This is a choice that seems reasonable to most users, though there are many users that would prefer a stronger focus on either the category as a whole ("i like all sports") or on its particular topics ("I don't care about football, but like Messi and Barcelona").
- **Persistence**. The balance between stable long-term user interests and short-term news context is delicate. If long-term interests are preferred, the user risks that there is no relevant news available and she will leave the app. Emphasizing the running context too much is also unfortunate, as this will make the app appear similar from one user to another.
- **Serendipity**. There is also a balance between profile-relevant news stories and serendipitous news. Most content providers like to include also news that are not directly within a user's profile to trigger new interests and widen her perspective (Das et. al, 2007). This is in our case achieved using content-based filtering for profile-driven news and collaborative filtering for serendipitous news, and the weighting of the techniques decide the balance between these two kinds of news.

In sum, it is difficult to identify an ideal user profile in news recommender systems. Users have different priorities than content providers, and the various weighting schemes that may be used to address the granularity, persistence and serendipity of news typically depend on whether they reflect the users' or the content providers' perspective.

It is also worth noting that freshness and location seem very important to users, even if the stories themselves are not necessarily the best match with their user profiles. In this respect news recommendation differs fundamentally from product recommendation on online shopping sites, where users have a clear and stable interest that calls for all relevant products independently of age or origin.

Our profile construction approach is in many ways similar to what is used in SCENE (Li et al., 2011). However, whereas we combine content phrases and entities in one vector, they split it up and also consider more complex phrases as part of their content description. A more fundamental difference, though, is the granularity of the user click analysis itself. As opposed to SCENE and most other news apps that look at news stories as either read or not read, we consider in more detail what the user is doing with the stories and use these smaller acts as cues for refining the users' real interests and preferences. This technique is to some extent also exploited in DailyLearner (Billsus and Pazzani, 2000), though the SmartMedia app takes every user gesture as an indication of user satisfaction or dissatisfaction.

7 CONCLUSIONS

The Smartmedia mobile news recommender system demonstrates both the value and serious challenges of implicit user profiling. Our mobile news recommender system uses extensive user profiles to recommend news stories that deal with the same topics as the profiles. However, the system is highly configurable, with a number of parameters that seriously affect the news stories recommended to the users. As there is no obvious best configuration of the system, since users and content providers have different agendas, only time will tell what is an acceptable weighting scheme for both readers and media houses.

The news app has been available online since the summer of 2013, and we are now in the process of building a new mobile user interface based on feedback from the early months. We are also – together with the media industry – collecting a large training set with real news and real users over several months, which we intend to use to refine our weighting schemes for balancing different recommendation strategies.

Research-wise we are gradually expanding the recommender system with semantic features for modeling news events and the entities involved in these events. This involves the definition of news taxonomies not very different from what is used in semantic search solutions (see e.g. Brasethvik & Gulla, 2002). With a new log-in feature we also intend to use social media sites like Twitter to expand and deepen the understanding of users' interests and preferences (e.g. O'Banion, 2012).

ACKNOWLEDGEMENTS

This research was supported by Telenor Group as part of their collaboration with the Department of

Computer and Information Science at the Norwegian University of Science and Technology in Trondheim.

REFERENCES

Billsus, D. and Pazzani,M. J., 2000. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10, pp. 147-180.

Borges, H. L. and Lorena, A. C., 2010. A Survey of Recommender Systems for News Data. In Szczerbicki & Nguyen (eds.), *Smart Information and Knowledge Management*, SCI 260, pp. 129-151. Springer.

Brasethvik, T. and Gulla, J. A., 2002. A conceptual modeling approach to semantic document retrieval. In Proceedings of the 14th international Conference on *Advanced Information Systems Engineering (CAiSE'02)*, pp. 167-182. Springer.

Cantador, I. Bellogin, A. and Castells, P., 2008. Ontology-Based Personalised and Context-Aware Recommendations of News Items. In Proceedings of the 7th International Conference on Web Intelligence, pp. 562-565. IEEE.

Das, A. S. Datar, M. Garg, A. and Rajaram, S., 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web,* pp. 271-280. ACM.

Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A., 2007. User profiles for personalized information access. *The adaptive web*, pp. 54-89. Springer.

Gulla, J. A. Auran, P. G. and Risvik, K. M., 2002. Linguistic Techniques in Large-Scale Search Engines. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems (NLDB'02)*, pp. 218-222.

Gulla, J. A., Ingvaldsen, J. E., Fidjestøl, A. D., Nilsen, J. E., Haugen, K. R., Su, X., 2014. Learning User Profiles in Mobile News Recommendation. Accepted for publication in *Journal of Print and Media Technology Research.*

Haugen, K. R., 2013. *Mobile News: Design, User Experience and Recommendation*. MSc thesis. NTNU, Trondheim.

Jannach, D. Zanker, M. Felfernig, A. and Friedrich, G., 2010. *Recommender Systems: An Introduction.* Cambridge University Press.

Kim, H. R. and Chan, P. K., 2003. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 101–108. ACM.

Li, L., Wang, D., Li, T., Know, D., and Padmanabhan, B., 2011. SCENE: a scalable two-stage personalized news recommendation system. In *Proceedings of SIGIR'11*, pp. 125-134. ACM.

Liu, J. Dolan, P. and Pedersen, E. R., 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on intelligent user interfaces*, pp. 31–40. ACM.

Lops, P., de Gemmis, M. and Semeraro, G., 2011. Content-based Recommender Systems: State of the Art and Trends. In Ricci, Rokach, Shapira and Kantor (Eds.), *Recommender Systems Handbook*, Chapter 3, pp. 73-106. Springer.

Nilsen, J. E., 2013. Large-Scale User Click Analysis in News Recommendation. MSc thesis, NTNU, Trondheim.

O'Banion, S. Birnbaum, L. and Hammond, K., 2012. Social media-driven news personalization. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*. pp. 45-52. ACM.

Rajaraman, A. and Ullman, J. D., 2011. *Mining of Massive Datasets.* Cambridge University Press.

Singh, S., Shepherd, M., Duffy, J. and Watters, C., 2006. An Adaptive User Profile for Filtering News Based on a User Interest Hierarchy. In Proceedings of the American Society for Information Science and Technology, Volume 43, Issue 1, pp. 1-21, 2006.

Solskinnsbakk, G. and Gulla, J. A., 2010. Combining ontological profiles with context in information retrieval. *Data & Knowledge Engineering*, 69(3), pp. 251-260.

Tavakolifard, M. Gulla, J. A. Almeroth, K. C. Ingvaldsen, J. E. Nygreen, G. and Berg, E., 2013. Tailored News in the Palm of your HAND: A Multi-Perspective Transparent Approach to News Recommendation. In *Proceedings of 22nd International World Wide Web Conference (WWW'13), Companion Volutme*, pp. 305-308, May, Rio de Janeiro.

Data Cube Computational Model with Hadoop MapReduce

Bo Wang[1], Hao Gui[1], Mark Roantree[2] and Martin F. O'Connor[2]
[1]*International School of Software, Wuhan University, Wuhan, China*
[2]*Insight: Centre for Data Analytics, Dublin City University, Dublin, Ireland*
{*bobwang, hgui*}*@whu.edu.cn,* {*mark.roantree, moconnor*}*@computing.dcu.ie*

Keywords: XML, Hadoop, Data Warehouse, MapReduce.

Abstract: XML has become a widely used and well structured data format for digital document handling and message transmission. To find useful knowledge in XML data, data warehouse and OLAP applications aimed at providing supports for decision making should be developed. Apache Hadoop is an open source cloud computing framework that provides a distributed file system for large scale data processing. In this paper, we discuss an XML data cube model which offers us the complete views to observe XML data, and present a basic algorithm to implement its building process on Hadoop. To improve the efficiency, an optimized algorithm more suitable for this kind of XML data is also proposed. The experimental results given in the paper prove the effectiveness of our optimization strategies.

1 INTRODUCTION

XML is a widely used and well structured data format especially in web-based information systems. Its flexible nature makes it possible to represent many kinds of data. The Web constantly offers new services and generates large volumes of new data in XML. However, it is difficult to satisfy the additional processing requirements necessary to facilitate OLAP applications or data mining with XML data. The underlying differences between the relational and XML data models present many challenges. It is difficult to provide a logical and direct mapping from one data model to the other due to the impedance mismatch between them (Rusu et al., 2009). In our previous works (Gui and Roantree, 2012) (Gui and Roantree, 2013), we have proposed a pipeline design based on an OLAP data cube construction framework designed for real time web generated sensor data. We transformed sensor data into an XML stream conforming to the data warehouse logical model and built a corresponding data cube tree and serialized it into an XML data cube representation.

In this paper, our research focuses on how to process large scale XML data efficiently. The concept of "cloud computing" has received considerable attention recently because it facilitates a solution to the increasing data demands through a shared and distributed computing infrastructure (Dutta et al., 2011). Apache Hadoop provides a powerful tool for tackling large-scale data problems in the area of machine learning, text processing, bioinformatics, etc. Hadoop implements a computational paradigm called MapReduce. The application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system to store data and permits a very high throughput for aggregate operations across the node cluster. MapReduce has emerged as an attractive alternative: its functional abstraction provides an easy-to-understand model for designing scalable and distributed algorithms (Lin and Schatz, 2010). Recently there has been some research into the provision of a parallel processing computational model for XML documents over distributed systems (Dede et al., 2011). In (Khatchadourian et al., 2011), the authors present a language called ChuQL to express XML oriented data processing tasks on the cloud. XML is a semi-structured data format, and due to its distributed paradigm, Hadoop is well positioned to provide a reliable and scalable platform for processing semi-structured data. By transforming large-scale XML data into the data cube presented in this paper, it will be easier to process the data on Hadoop and significantly reduce the risk of data loss.

2 BACKGROUND

2.1 CityBike Project

The deployment of sensors in the physical world is constantly increasing and may now be regarded as widespread. The number of applications relying on sensor data is also growing. Examples include: urban traffic watch, weather monitoring, tracking of goods, etc. We now describe a project that serves as a use case for our work. The city of Dublin (along with many other European cities) has deployed a bike sharing scheme whereby people may rent (and return) bikes from stations located throughout the city center. The stations are equipped with sensors to monitor the availability of bikes, and the stations publish this information to the DublinBikes website (www.dublinbikes.ie). Consumers can connect to the website (either through a desktop PC or mobile application) to view this information, including the location of stations, the number of bikes currently available for renting, the number of spaces available to return bikes, and so on. For both consumers and providers of the service, the data is of great interest. Consumers can check where to rent or return a bike while providers can understand at which station it is best to pick up or return bikes for maintenance in order to minimize service disruption.

The bicycle rental application, available in many cities and towns across Europe, collects data at regular intervals from each location. For each station this information is made available through the service provider's web site which provides the station ID, the total number of bike stands, the number of bikes available, the number of free bike stands available, and so on. The bicycle rental statistics may also be connected with other related factors, such as weather conditions, bus routes in the city, and the location of local train stations.

2.2 XML Data Cube Construction

The Data Warehouse (DW) definition is different from the schema of the underlying native XML database. The DW definition includes a high level description of all the dimensions and the fact data and the relationships between them (as shown in Fig.1). In contrast, the schema of the XML database focuses on the low-level logical and physical structure of the XML data. The initial web generated raw sensor data must be transformed and mapped to the structure described by our DW definition illustrated in Fig.1 It is not necessary to permanently materialize the identifier information in order to build up the XML DW

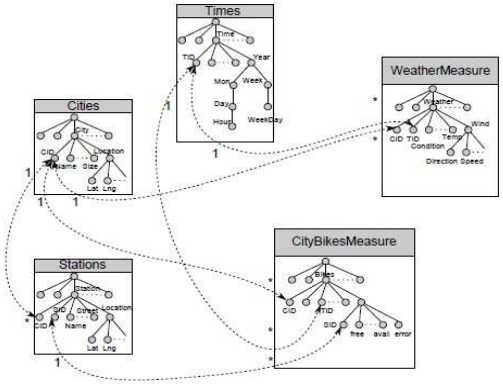


Figure 1: An example of logical model for the data warehouse in CityBikes Project (Gui and Roantree, 2013).

galaxy model and to specify the relationships between the different parts. (Gui and Roantree, 2013) The cube definition details the key components of the cube construction process: selected dimensions, selected concept hierarchy within dimensions, measurements of fact data and aggregating functions. This information can be used to determine both the construction process and the production of the data cube. In our project, we have developed several XML Schemata to model the data cube structure, which we name the XML Data Cube Model (XDCM). The data cube can be serialized into XML data and an associated index may be generated to facilitate OLAP operations on the data cube. Both the serialized XML data cube and the corresponding index (not discussed in this paper) may be stored into the underlying native XML database and be made available to related data analysis applications.

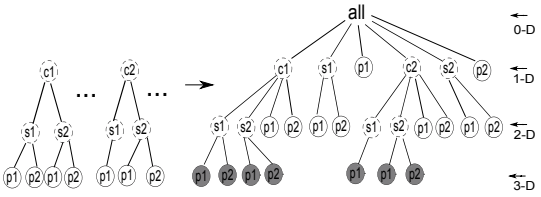


Figure 2: XDCM cube tree construction demonstration.

In Fig.2, the forest on the left side is a simple illustration of the input XML data concerning bikes usage, which contains source information to be aggregated. To simplify the example, we only selected three dimensions and each dimension has no more than two different values. In the input forest, every node at each level may be repeated an arbitrary number of times. In the CityBike project, the system generates the data regularly, and in some dimensions, the domains are quite limited regarding the number of stations and the weather details. This will result in a large quantity of mergers in subsequent operations. The height of

input tree is equivalent to the dimension in the DW. Assuming the dimension of input data is D, then we will get 2^D views to observe the data, including dimension 0 . Our XDCM must be robust enough to describe all the views directly without any join or related operations. In the case of the forest on the left side in Fig.2, using the capital letters to represent the dimensions, we could get $C-S-P$, $C-P$, $C-S$, $S-P$, C, S, P and Φ. In fact, $C-S-P$ includes $C-S$ and C, $S-P$ includes S. From a tree's point of view, you must visit the nodes from top to bottom along the paths. If you have reached P, it follows that you have already reached C and S. To reduce the redundancy, some views that are included by others will not be generated directly in the XDCM. In other words, all the branches of XDCM cube tree should end with the nodes in P in this demonstration. The Φ view is the "all" in the XDCM cube tree at level 0.

The tree on the right side is the XDCM tree be be constructed. There are three kinds of nodes that will be used to finalise the construction.

- **Data Node.** The gray nodes in Fig.2 represent the Data Nodes. After several processes such as filtration or combination (if required), the initial data can be generated on the Data Nodes directly. The Data Nodes contain basic information such as paths and values, and they describe the view with the longest path such as $C-S-P$ highlighted in the previous example.
- **Additional Node.** The Additional Node (indicated by a dashed-circle in Fig.2) which is produced by the Data Node is a form of intermediate data not shown in Fig.2. Each Data Node will produce all possible Additional Nodes through the removal of any constraints in the different dimensions in its path.
- **Link Node.** The white nodes represent the Link Nodes which describe other additions paths of observation. Link Nodes are produced by combining Additional Nodes that share the same paths. Nodes in different paths should be unique in the XDCM.

During the cube construction, for example, Data Node $all-c1-s1-p1$ will produce Additional Nodes $all-c1-p1$, $all-s1-p1$, $all-p1$, and Data Node $all-c1-s2-p1$ will produce Additional Nodes $all-c1-p1$, $all-s2-p1$, $all-p1$. Thus, there are two pairs of nodes that share the same paths. They are $all-s1-p1$ and $all-p1$. Upon combination of these nodes, we obtain Link Nodes $all-c1-p1$, $all-s1-p1$, $all-s2-p1$ and $all-p1$. Both Data Nodes and Link Nodes are the nodes in P. Other nodes describe other views can be calculated in following processes.

2.3 Map-Reduce

MapReduce is a programming framework for processing large data sets with a parallel, distributed algorithm on clusters. The most important functions in this model are Map and Reduce. Mappers receive a collection of Key-Value pairs and produce zero or more output Key-Value pairs. Pairs sharing the same key are collected and delivered to the same Reducer. Reducers can iterate through the values that are associated with that key and produce zero or more outputs. Hadoop has implemented this programming model and offers a well-defined set of APIs to control the entire process.

In Hadoop, a complete Map-Reduce task can be regarded as a job. To finish an entire task, it may be necessary to perform more than one job. Input data is split into many parts and transformed into Key-Value pairs. In this context, the concept of *split* is about a logical rather than a physical operation. Consequently, the data uploaded to the Hadoop Distributed File System (HDFS) will be cut into several blocks, usually in 64MB chunks, according to the requirements of the system. Hadoop is quite adept at dealing with unstructured data, such as text, because it is easy to split and suitable for the physical structure. Hadoop does not offer any functionality to process XML documents directly. Mahout, an open-source program based on Hadoop, provides a class to handle the XML input format, and the underlying strategy is simple. An XML document is well organized, the content of every element is enclosed by a pair of tags. Therefore, it is sufficient to simply read the XML document as a text file and select the content between the certain pair of tags.

3 CUBE CONSTRUCTION WITH MAP-REDUCE

In order to describe the algorithm, consider an exercise to build an OLAP data cube for the bicycle rental scenario under certain weather conditions and in various locations and dates. In this case, only three dimensions have been chosen as shown in Fig.2. The Input data are nodes in the dimension named P with some additional information, such as their path constraints. On the left side, the nodes indicated with a dashed-circle illustrate the paths of the nodes circled by solid lines, and they do not need to be instantiated. Our goal is to build the tree on the right side and to calculate the values of the dashed nodes.

For our implementation, the entire task can be divided into two phases. Phase one is to generate the Data Nodes and Link Nodes, and then organize them into the XDCM as described earlier. The second phase is to calculate the values of the dashed nodes on the right side which represent observation views other than Data Nodes and Link Nodes. The calculation task itself can be divided into two further subtasks. One subtask is used for the calculation of Data Nodes and the other subtask is used for the calculation of Link Nodes.

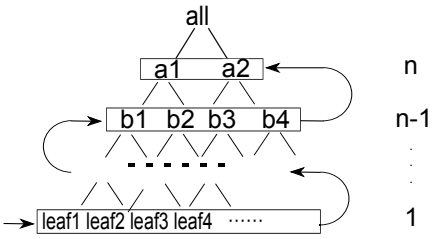


Figure 3: Processing Order of Data Nodes.

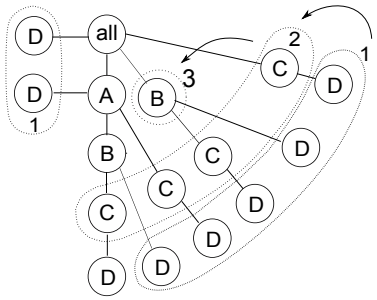


Figure 4: Processing Order of Link Nodes.

Fig.3 illustrates the node processing order of the algorithm. The integer values on the right side represent the round indexes of Map-Reduce jobs. Data Nodes are the leaf nodes shown at the bottom, and they will be processed by the first job to generate Additional Nodes. The Additional Nodes will subsequently be merged into Link Nodes. Thus, the subsequent rounds only concentrate on the calculation tasks. For a N-dimensional XML data, we need at least N rounds to complete the entire construction, in other words, N jobs are needed to complete the whole task in Hadoop. The output generated from each job will be used as the input data for the next job. The computation receives a set of Key-Value pairs as input, and then produces a set of output Key-Value pairs:

$$Map \quad (k1, v1) \quad \rightarrow \quad list(k2, v2)$$
$$Reduce \quad (k2, list(v2)) \quad \rightarrow \quad (k3, v3)$$

In the generation phase (phase one), the initial data is transformed into Key-Value pairs that actually represent Data Nodes. So $k1$ is the path of a given Data Node, and $v1$ is the value of that node. Data Nodes will produce large amounts of Additional Nodes. $k2$ is the path of an Additional Node, and $v2$ is its value. The Reduce function accepts the intermediate key $k2$ and a set of values for that key, and merges these values to form a Link Node. So $k3$ is the path of Link Node, and $v3$ is the actual value.

In the calculation phase (phase two), the process starts with either the Data Nodes or the Link Nodes. All nodes in the same level will be processed in one job. The entire process proceeds in a bottom-up manner from one layer to the next. However, unlike the generation phase, the Key value of the Mapper's output is the path of the node without self-containment. For example, given the input node γ and the path $\alpha-\beta-\gamma$. Then $k2$ will be $\alpha-\beta$, and $v2$ will be γ. The Reduce function accepts the intermediate key $\alpha-\beta$ and calculates the value of β. The $k3$ is the complete path of β which is $\alpha-\beta$, and $v3$ is β. This $(k3, v3)$ pair will be the $(k1, v1)$ in the next calculation round, which is going to calculate the value of α.

Given that all Link Nodes are derived from Data Nodes and all nodes in the XDCM are unique, some ancestor nodes may be calculated using different descendants in different paths. For example, in Fig.4, the node B in $all-B-C-D$ may be calculated by $all-B-D$ also. To avoid repeated calculations, we employ a "BreakPoint" to indicate the point at which to stop the calculation. The procedure for selecting a BreakPoint in the basic implementation is as follows:

1. The BreakPoint represents a node in a path. If the BreakPoint is going to be calculated in the next round, the calculation should be stopped.
2. Only Link Nodes and their ancestor nodes need to keep a record of BreakPoints. All nodes in the path of Data Nodes will be calculated. So it is unnecessary to keep records of BreakPoints for Data Nodes and their ancestor nodes.
3. BreakPoints are identified in the generation phase and won't change in the following calculations. The ancestor nodes of each Link Node should inherit its BreakPoint.
4. It should be noted that BreakPoints are the places where new branches are added while building the XDCM cube tree. For instance, in Fig.4, all the branches except $all-A-B-C-D$ are branches that were added. As for $all-A-C-D$, the BreakPoints are the nodes at A, because $-C-D$ is the branch we added.

Fig.4 shows the abstract structure of XDCM in four dimensions. As described in section 2.2, the paths of all nodes start from all and end at D. Thus, the total number of Link Nodes is the combinations

from A to C minus one (because we already have the longest path, $A-B-C$). For the N-dimensional data, the number of Additional Nodes for each Data Node is $C_{N-1}^{N-2}+C_{N-1}^{N-3}+...+C_{N-1}^{0}(N\geq 2)$.

In our implementation, we use binary numbers to describe the status of those combinations. In Fig.4, the number of Additional Nodes for each Data Node is $C_3^2+C_3^1+C_3^0=7$. So using three bits $(2^{(3-1)})$ can describe all cases. Each bit stands for a dimension between *all* to *D*. If a certain bit is set to 1, it indicates having the constraint on that dimension. To identify a BreakPoint for each Link Node, we need to check whether the prefix of the Link Node's path has already existed in the tree that we have currently built. we may enumerate all of the cases by reducing the binary number from the maximum to zero. We determine the BreakPoint by comparing the prefix of two adjacent cases. The maximum of the binary number should compare with the binary number that stands for the Data Node. In the previous example, the 110 should compare with 111, so the BreakPoint for 110 is at the second bit which represents the *B* dimension. If there is no common substring in the prefix (as with 100 and 011), the BreakPoint for 011 is the node *all*.

4 OPTIMIZATION AND PERFORMANCE

4.1 Configuration Optimization

We performed the experiments in a cluster with 35 slave nodes each containing a 3.10GHz processor, 1 GB of RAM, and 40 GB of local disk allocated to the HDFS. Each node is running Hadoop version 1.1.2 on Red Hat 9.0 and connected by Fast Ethernet to a commodity router.

An XML document contains more information than an unstructured document with equivalent content. In order to handle XML documents in a distributed system without loss of information we need to restructure the data and divide it into small units. Thus, the amount of data through the system is larger than the input data. The table in Fig.5 shows the size of the data and the number of records through the system in the first round.

Each record in Hadoop represents a unit we made which is described as a node in section two and three. In the experiments, we use two types of XML files for the experiments, of three-dimensions and six-dimension respectively. The arity of each dimension is nine. For 1 GB XML file in three dimensions, using the basic algorithm, the Map output in the first job is about 4.6 GB. In other words, to build the XDCM directly, the size of data we actually need is approximately 4.6 GB. For 1 GB XML file in six dimensions, using the basic algorithm, the Map output in the first job reaches 41 GB. Since Mappers and Reducers are separated, these outputs are transferred through the network. The I/O operations may have a negative impact on the whole task.

	Dimension	Map input bytes	Map output bytes	Map output records	Reduce input records	Reduce input records(after combining)
1	3	1, 023, 266, 669	4, 973, 220, 180	126, 360, 000	126, 360, 000	35, 101, 440
2	3	5, 386, 519, 202	24, 703, 200, 180	622, 080, 000	622, 080, 000	180, 001, 136
3	6	1, 104, 659, 389	44, 683, 382, 133	1, 075, 164, 192	1, 075, 164, 192	51, 916, 510
4	6	5, 514, 648, 663	130, 268, 589, 390	2, 686, 965, 696	2, 686, 965, 696	250, 036, 747

Figure 5: comparing the data through the system in the first round.

Hadoop offers several ways to optimize the I/O. In the experiments, we mainly used compression and combination, and their effects were clear. Hadoop supports several compression formats like gzip and snappy. The compression method should be selected according to the type of the task to be performed. If the task's CPU occupancy rate is high, it is better to choose a simple compression algorithm. If a significant portion of the task is spent at I/O, it is better to select an algorithm with a high compression ratio. in our experiment, using the default compression algorithm offered by Hadoop to process 1 GB of XML data in three dimensions provides approximately an 8% improvement.

In our experiments, the combination process provided a demonstrable improvement. The task of a Combiner is similar to that of a Reducer. If the Combiner is used then the outputs from the Mapper are not immediately delivered to the Reducer. The Key-Value pairs are collected in lists, one list per Key. The Combiner will process each list like a Reducer and emit a new Key-Value pair which has already been merged. Then the new Key-Value pair will be delivered to the Reducer as if they were created by the original Map operation. Fig.6 shows the performance using these two strategies when dealing with the three-dimensional XML files.

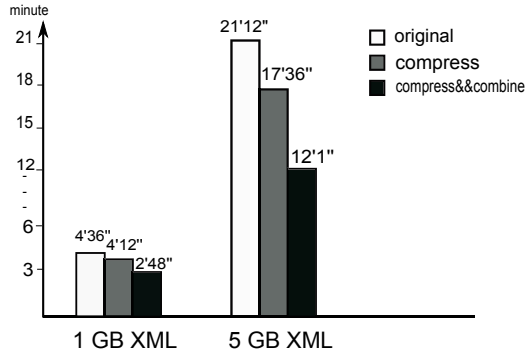


Figure 6: Performance of Optimization for 3-D XML.

4.2 Algorithm Optimization

For the 1 GB and three-dimensional XML file in our experiment, using the basic algorithm discussed in section 3, the number of Map output records is about 126,360,000 in the first round but the number of Reduce output records is approximately 35,101,440. About 72% records are merged in Reducers. For the 1 GB and six-dimensional XML file, the number of Map output records is approximately 1,075,164,192 but the number of Reduce output records is only 51,9156,510. The gap in the 5 GB file sizes is even more pronounced. As the dimension increases, the output of Map increases more rapidly. For a D-dimensional XML file, assuming n is the number of initial Data Nodes, the direct output records of Map is $2^D \times n$. In the basic algorithm, the workload is too large when the dimension becomes higher in the first round. Too much data is transferred from the Mapper to the Reducer. In a cloud environment with limited capacity, it may increase the risk of the jobs failing.

In the basic algorithm, all Additional Nodes come from Data Nodes. However, some nodes can be generated by Link Nodes also. The Link Nodes with longer paths can produce the Additional Nodes with shorter paths. In Fig.4, Nodes in $all-D$ can not only be generated by $all-A-B-C-D$ (Data Nodes), but also can be generated by other six types of Link Nodes like $all-A-B-D$, $all-B-D$, etc. Although the information in $all-A-B-D$ or $all-B-D$ is not as complete as the Data Nodes, it is enough for $all-D$. The significant advantage of using shorter-path nodes to generate Link Nodes is the decrease in the I/O between Mapper and Reducer. For example, there are only two Data Nodes, the paths of which are $all-a1-b1-c1-d1$ and $all-a2-b1-c1-d1$. Thus, the Link Node in $all-B-D$ will be $all-b1-d1$. It is more efficient to generate the Link Node in $all-D$ by using Link Node $all-b1-d1$ instead of using the Data Nodes because only one output record is produced by the Mapper, whereas the Data Nodes would produce two.

Fig.7 is an example of generating Link Nodes by using an optimized algorithm for five-dimensional data. As discussed in section 3, all of the possible Link Nodes for five-dimensional Data Nodes ($all-A-B-C-D-E$) depend on combinations of the path between all and E exclusively, which is $A-B-C-D$. In the first round, by removing one constraint in the path $A-B-C-D$, we get C_4^1 kinds of Link Nodes shown in column 1 in Fig.7. The arrow above the node is the indicator of a BreakPoint. The arrow pointing to empty means the BreakPoint is at all. Any Link Node whose BreakPoint does not point

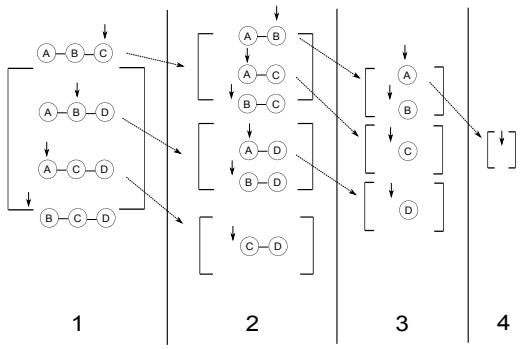
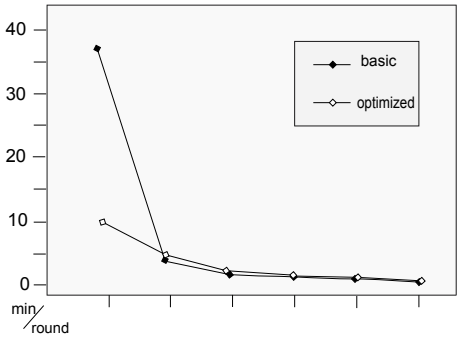


Figure 7: Using Link Nodes for Generation.

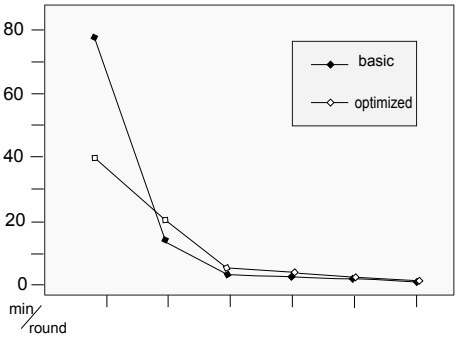
to all can be used to generate new Link Nodes in the next round. In the next round, another constraint of the Link Nodes is to be removed, and this constraint should be the nodes before the BreakPoint. As shown in Fig.7, the $A-B-C$ in column 1 stands for the Link Nodes whose path is $all-A-B-C-E$. The BreakPoint is at C, so in next round shown in column 2, successively removing the nodes before C (including C), which are C, B, A, we obtain $A-B$, $A-C$, $B-C$. The BreakPoint changes to the point at the node before the node has been removed. If the BreakPoint points at all like $B-C$, it would not participate in the generation work of future rounds, but the calculation phase would continue. In this way, we obtain all of the combinations after the removal of two constraints from the path, which is C_4^2 in total. The Link Nodes are generated according to the rules described above, until all BreakPoints point to all. The total number is $C_4^1+C_4^2+C_4^3+C_4^4=15$, which is the same number produced by the basic algorithm.

Assuming the dimension is D, in the Nth round, the number of Additional nodes for each Link Node is C_{D-1}^N. It would appear that the I/O should increase when N is close to $\frac{D-1}{2}$. However, it does not because the Additional Nodes are produced by Link Nodes instead of Data Nodes, and after several rounds of mergers, the number of Link Nodes decrease significantly. The total outputs from the Mapper after N rounds are still less than the first round in our experiment.

Fig.8 shows the different efficiencies obtained by employing these two algorithms to construct the XDCM using the six-dimensional XML data. The improvement in the performance is significant. For the 1 GB and 6-dimensional XML file, the total improvement is approximately 64%. For the 5 GB and 6-dimensional XML file, the improvement is approximately 27%. Fig.9 shows the the performance of the two algorithms using the 5GB XML data from CityBikes.



1 GB XML in 6 dimensions



5 GB XML in 6 dimensions

Figure 8: Basic Algorithm vs. Optimized Algorithm.

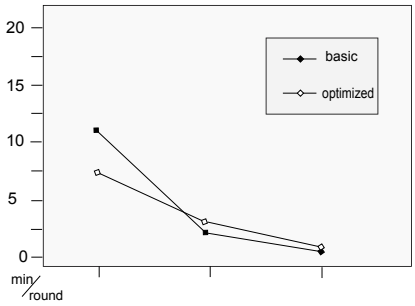


Figure 9: Performance in CityBike.

5 CONCLUSIONS

In this paper, we illustrated a data cube model for XML documents to meet the increasing demand for analyzing massive XML data in OLAP. Hadoop is a popular framework aiming at tackling large-scale data problems. We proposed a basic algorithm to construct the XDCM on Hadoop. To improve efficiency, we offered some strategies and described an optimized algorithm. The result proves the optimized algorithm is suitable for this type of data and further enhance the efficiency.

REFERENCES

Dede, E., Fadika, Z., Gupta, C., and Govindaraju, M. (2011). Scalable and distributed processing of scientific xml data. In *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*, pages 121–128. IEEE Computer Society.

Dutta, H., Kamil, A., Pooleery, M., Sethumadhavan, S., and Demme, J. (2011). Distributed storage of large-scale multidimensional electroencephalogram data using hadoop and hbase. In *Grid and Cloud Database Management*, pages 331–347. Springer.

Gui, H. and Roantree, M. (2012). A data cube model for analysis of high volumes of ambient data. *Procedia Computer Science*, 10:94–101.

Gui, H. and Roantree, M. (2013). Using a pipeline approach to build data cube for large xml data streams. In *Database Systems for Advanced Applications*, pages 59–73. Springer.

Khatchadourian, S., Consens, M. P., and Siméon, J. (2011). Having a chuql at xml on the cloud. In *AMW*.

Lin, J. and Schatz, M. (2010). Design patterns for efficient graph algorithms in mapreduce. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 78–85. ACM.

Rusu, L. I., Rahayu, W., and Taniar, D. (2009). Partitioning methods for multi-version xml data warehouses. *Distributed and Parallel Databases*, 25(1-2):47–69.

Distributed XML Processing over Multicore Servers

Yoshiyuki Uratani[1], Hiroshi Koide[2] and Dirceu Cavendish[3]
[1]*Global Scientific Information and Computing Center, Tokyo Institute of Technology, O-okayama 2-12-1, Meguroku, Tokyo, 152-8550, Japan*
[2]*Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Kawazu 680-4, Iizuka, Fukuoka, 820-8502, Japan*
[3]*Network Design Research Center, Kyushu Institute of Technology, Kawazu 680-4, Iizuka, Fukuoka, 820-8502, Japan*
uratani.y.ab@m.titech.ac.jp, koide@ai.kyutech.ac.jp, cavendish@ndrc.kyutech.ac.jp

Keywords: Distributed XML Processing, Task Scheduling, Pipelining and Parallel Processing, Multicore CPU.

Abstract: Nowadays, multicore CPU become popular technology to enhance services quality in Web services. This paper characterizes parallel distributed XML processing which can off-load the amount of processing at their servers to networking nodes with varying number of CPU cores. Our implemented distributed XML processing system sends XML documents from a sender node to a server node through relay nodes, which process the documents before arriving at the server. When the relay nodes are connected in tandem, the XML documents are processed in a pipelining manner. When the relay nodes are connected in parallel, the XML documents are processed in a parallel fashion. For well-formedness and grammar validation tasks, the parallel processing reveals inherent advantages compared with pipeline processing regardless of document type, number of CPU cores and processing environment. Moreover, the number of CPU cores impacts efficiency of distributed XML processing via buffer access contention.

1 INTRODUCTION

Web services become necessary infrastructure for our society. Various services are being provided and they need much resources, such as CPU power and memory spaces, to enhance their services quality. Nowadays, multicore CPU is popular approach to improve processing capacity. The processing for Web services are generally provided at only server nodes in the current Web services. In this situation, we propose offloading approach, which assign a part of server's load to intermediate nodes in network, to reduce the load and to improve services throughput. This approach will lead efficient resource usage with effective scheduling and higher quality of services. On the other hand, XML data is one of the basic communication format in the Web services and the servers often processes the XML data in various situation (e.g. collaborative services). PASS-Node (Cavendish and Candan, 2008) had already proposed distributed XML processing with using intermediate nodes and had studied from an algorithmic point of view for well-formedness, grammar validation, and filtering. We have also focuses on distributed XML processing with offloading approach and have studied their processing characteristics (Uratani et al., 2012).

This paper augments practical distributed XML processing characterization scope by varying the number of CPU cores in processing nodes; using synthetic and realistic XML documents; on pipelining model for XML data stream processing systems, on parallel model for parallel processing system. Regarding processing efficiency, we investigate XML processing performance relation to XML document characteristics,as well as the impact of number of CPU cores to well-formedness and grammar validation tasks. Our results can be summarized as follows. Parallel processing performs better than pipeline processing regardless the number of CPU cores, document types and processing environment; Processing with more CPU cores leads to faster processing; but it also includes drawbacks, such as inefficient resource consumption, due to buffer contention.

The paper is organized as follows. In section 2, we describe generic models of XML processing nodes. In section 3, we describe experimental environments and characterize XML processing performance of the pipeline and parallel computation models on multicore machines. In section 4, we address related work. In section 5, we summarize our findings and address research directions.

2 XML PROCESSING ELEMENTS

Distributed XML processing requires some basic functions to be supported: **Document Partition:** The XML document is divided into fragments, to be processed at processing nodes. **Document Annotation:** Each document fragment is annotated with current processing status upon leaving a processing node. **Document Merging:** Document fragments are merged so as to preserve the original document structure. XML processing nodes support some of these tasks, according to their role in the distributed XML system.

2.1 XML Processing Nodes

We abstract the distributed XML processing elements into four types of nodes: StartNode, RelayNode, EndNode, and MergeNode. The distributed XML processing can then be constructed by connecting these nodes in specific topologies, such as pipelining and parallel topologies. Distributed processing applications may be constructed with both parallel and pipeline manner. So the study of pipelined approach is to opportunistically process XML documents on arbitrary network topologies, in the middle of the network.

`StartNode(SN)` is a source node that executes any pre-processing needed in preparation for piecewise processing of the XML document. This node also starts XML document transfer to relay nodes and adds some annotation which are used for distributed XML processing. The StartNode has multiple threads for reading part of the document and sending fragments simultaneously.

`RelayNode(RN)` executes XML processing on parts of an XML document. It is placed as an intermediate node in paths between the StartNode and the EndNode. The RelayNode has three types of threads: ReceiveThread, TagCheckThread and SendThread. The ReceiveThread receives data containing lines of an XML document, together with checking and processing information, and stores the data into a shared buffer. The TagCheckThread attempts to process the data, if the data is assigned to be processed at the node. SendThread sequentially sends data to a next node.

EndNode(EN) is a destination node, where XML documents must reach, and have their XML processing finished. This node receives data containing the XML document, checking information and processing information, from a previous node. If the tag checking has not been finished yet, the EndNode processes all unchecked tags, in order to complete XML processing of the entire document. Components of the EndNode are similar to the RelayNode, except that the EndNode has DeleteThread instead of SendThread. The DeleteThread cleans the document from processing and checking information.

`MergeNode(MN)` receives data from multiple previous nodes, serializes it, and sends it to a next node, without performing any XML processing. MergeNode has multiple threads for receiving data from previous node, so as not to block previous nodes from sending data.

XML document processing involves stack data structures for tag processing. When a node reads a start tag, it pushes the tag name into a stack. When a node reads an end tag, it pops a top element from the stack, and compares the end tag name with the popped tag name. If both tag names are the same, the tags match. The XML document is well-formed when all tags match. In addition, in validation checking, each node executing grammar validation reads DTD files, and generates grammar rules for validation checking. Each node processes validation and well-formedness at the same time, comparing the popped/pushed tags against grammar rules. Details of these node distributed processing is described in (Cavendish and Candan, 2008; Uratani et al., 2012).

3 DISTRIBUTED XML CHARACTERIZATION

3.1 Experimental Environment

We use a **`VM_Env`**, which consists of a VMware ESX 4 on a Sun Fire X4640 Server, for providing distributed XML processing system. We use VMware ESX 4, a virtual machine manager, to implement a total of seven counts of virtual machines as distributed XML processing nodes. This environment consists of two types of virtual machine. We allocate two CPU cores to one of them (Node06), and four CPU cores to all

Table 1: X4640 Server Specification.

CPU	Six-Core AMD Opteron Processor 8435 (2.6GHz) × 8
Memory	256G bytes (DDR2/667 ECC registered DIMM)
VMM	VMware ESX 4
Guest OS	Fedora15_x86_64
JVM	Java™1.5.0_22

other nodes. The server specification is described in Table 1.

3.2 Node Allocation Patterns

We prepare several topologies and task allocation patterns to characterize distributed XML processing, within the parallel and pipelining models. We vary the number of RelayNodes, within topologies, as well as the number of CPU cores in processing nodes, to evaluate their impact on processing efficiency. To characterize the performance, we use a total of four topology types: two stage pipeline system (Figure 1), two path parallel system (Figure 2), four stage pipeline system, and four path parallel system. In the Figure 1 and 2, tasks are shown as light shaded boxes, underneath nodes allocated to process them. CPU core count is shown above the task boxes.

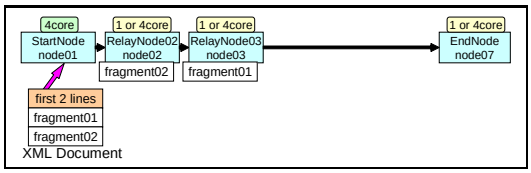


Figure 1: Two Stage Pipeline System.

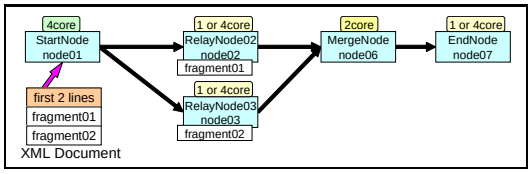


Figure 2: Two Path Parallel System.

For instance in two stage pipeline case (Figure 1), we divide the XML documents into three parts: first two lines (it contain a meta tag and a root tag), fragment01 and fragment02. These fragments are divided into segments of roughly the same size. Data flows from StartNode to EndNode via two RelayNodes. RelayNode02 is allocated for processing fragment02, RelayNode03 is allocated for processing fragment01, and the EndNode is allocated for processing the first two lines, as well as processing all left out unchecked data.

In Figure 2, we depict two path parallel system. The XML document is also divided into three parts. They flow from StartNode to EndNode via RelayNode01 and MergeNode. The StartNode reads concurrently these fragments from the XML document and sends them to the RelayNodes. Fragment02 and related data flow from the StartNode to the EndNode via RelayNode02 and the MergeNode. In addition, we can use a maximum of 4 CPU cores in node02, node03 and node07, which are allocated for the RelayNodes and EndNode. We apply two types of CPU usage patterns in this topology; i) Allocation of 1 CPU core only to RelayNodes and EndNode; ii) Allocation of all 4 CPU cores to RelayNodes and EndNode. These document partition and allocation patterns are defined beforehand as static task scheduling. We also experiment four RelayNode topology for both processing types with the five parted XML documents. In the processing types, we also vary number of CPU cores of the RelayNodes and the EndNode.

Notice that, even though the parallel model has one extra node, the MergeNode, as compared with corresponding pipeline model. However, the MergeNode does not perform any XML processing, so the number of nodes executing XML processing is still the same in both models.

3.3 Tasks and XML Document Types

The distributed XML processing system can execute two types of processing: well-formedness checking, and grammar validation checking of XML documents. Efficiency of these XML processing tasks may be related to: processing model, pipelining and parallel; topology, number of processing nodes and their connectivity; XML document characteristics. We use different structures of XML documents to investigate which distributed processing model yields the most efficient distributed XML processing. For that purpose, we create seven types of synthetic XML documents by changing the XML document depth from shallow to deep while keeping its size almost the same. We have also used three types of realistic XML documents, doc_kernel, doc_stock and doc_scala. The doc_kernel encodes directory and file hierarchy structure of linux kernel 2.6.39.3 in XML format. The doc_stock is XML formatted data from MySQL data base, containing a total of 10000 entries of dummy stock price data. The doc_scala is based on “The Scala Language Specification Version 2.8” (http://www.scala-lang.org/), which consists of 191 pages, totaling 1.3M bytes. The original document in pdf format was converted to Libre Office odt format, and from that to XML format. Their characteristics are shown in Table 2, For VM_Env, we combine four node allocation patterns, two processing patterns and ten XML document types to produce a total 80 types of experiments.

3.4 Performance Indicators

We use two types performance indicators: system performance indicators and node performance indicators. System performance indicators characterize the processing of a given XML document. Node perfor-

Table 2: XML Document Characteristics.

	doc01	doc02	doc03	doc04	doc05	doc06	doc07	kernel	stock	scala
Width	10000	5000	2500	100	4	2	1	-	-	-
Depth	1	2	4	100	2500	5000	10000	-	-	-
Tag set count (Empty tags)	10000 (0)							225 (36708)	66717 (146)	26738 (1206)
Line count	10002	15002	17502	19902	19998	20000	20001	41219	78010	72014
File size [Kbytes]	342	347	342	343	342			3891	2389	2959

mance indicators characterize XML processing at a given processing node. The following performance indicators are used to characterize distributed XML processing:

Job Execution Time is a system performance indicator that captures the time taken by an instance of XML document to get processed by the distributed XML system in its entirety. As several nodes are involved in the processing, the job execution time results to be the period of time between the last node (typically EndNode) finishes its processing, and the first node (typically the StartNode) starts its processing. The job execution time is measured for each XML document type and processing model.

Node Thread Working Time is a node performance indicator that captures the amount of time each thread of a node performs work. It does not include thread waiting time when blocked, such as data receiving wait time. It is defined as the total file reading time, data receiving time, data sending time and node buffer access time a node incurs. For instance, in the MergeNode, the node thread working time is the sum of receiving time and sending time of each Receive/SendThread. Assume the MergeNode is connected to two previous nodes, with two Receive/SendThreads. We also derive a **System Thread Working Time** as a system performance indicator, as the average of node thread working time indicators across all nodes of the system.

Node Active Time is a node performance indicator that captures the amount of time each node runs. The node active time is defined from the first ReceiveThread starts receiving first data until the last SendThread finishes sending last data in the RelayNode or finishes document processing in the EndNode. Hence, the node active time may contain waiting time (e.g, wait time for data receiving, thread blocking time). We also define **System Active Time** as a system performance indicator, by averaging the node active time of all nodes across the system.

Node Processing Time is a node performance indicator that captures the time taken by a node to execute XML processing only, excluding communication and processing overheads. We also define **System Processing Time** as a system performance indicator, by averaging node processing time across all nodes of the system.

Parallelism Efficiency Ratio is a system performance indicator defined as "*system thread working time* / *system active time*".

Node Buffer Access Time is a node performance indicator that captures the amount of time a node spends accessing internal shared buffers. As previously mentioned in section 2, threads in each node access the shared buffer while receiving/processing/sending data. The node buffer access time includes not only the time for add/get/remove operation, but also waiting time during blocking. We also define **System Buffer Access Time** as a system performance indicator, by totaling the node buffer access time of all nodes across the system.

3.5 Experimental Results

For each experiment type (scheduling allocation and distributed processing model), we collect performance indicators data over seven types of XML document instances. On all graphs, X axis describes scheduling, processing models and processing environment, for well-formedness and grammar validation types of XML document processing, encoded as follows: **`PIP_wel`**:Pipeline and Well-formedness checking, **`PAR_wel`**:Parallel and Well-formedness checking, **`PIP_val`**:Pipeline and Validation checking, **`PAR_val`**:Parallel and Validation checking. Y axis denotes specific performance indicator, averaged over 22 XML document instances. These figures are only part of the experimental results - details of other results are omitted for space's sake.

Regarding job execution time (Figures 3 – 6), parallel processing is faster than pipeline processing for all documents. Job execution time gets lengthened in pipeline processing due to the fact that the nodes relay extra data that is not processed locally. In addition, we can see that the job execution time speeds up faster with increasing number of relay nodes in

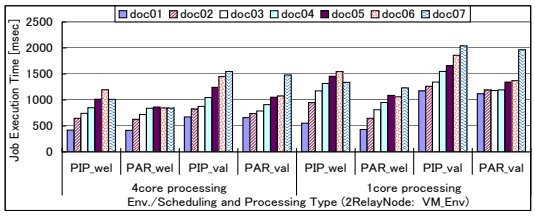


Figure 3: Job Execution Time (Syn. Doc. Proc. by 2 RN).

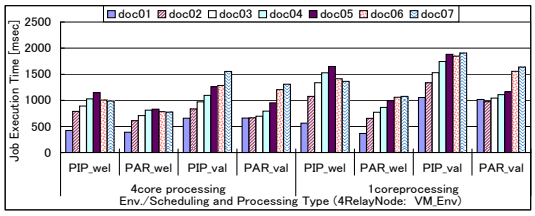


Figure 4: Job Execution Time (Syn. Doc Proc. by 4 RN).

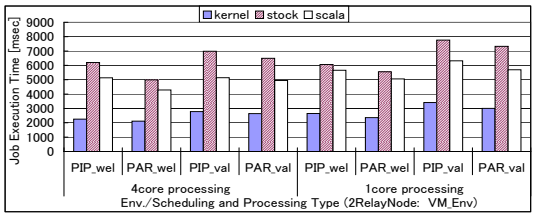


Figure 5: Job Execution Time (Real. Doc. Proc. by 2 RN).

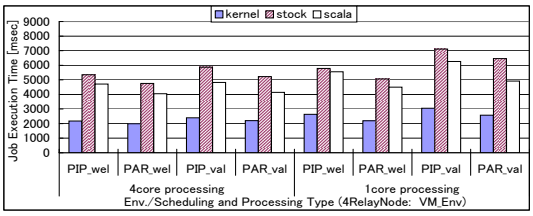


Figure 6: Job Execution Time (Real. Doc. Proc. by 4 RN).

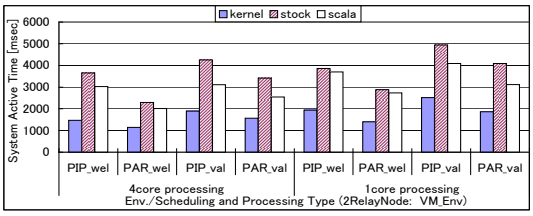


Figure 7: System Active Time (Real. Doc. Proc. by 2RN).

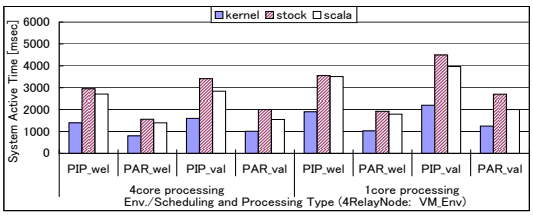


Figure 8: System Active Time (Real. Doc. Proc. by 4RN).

parallel processing than in pipeline processing. The extra data transfer time also appears in pipeline processing. Moreover, increased number of relay nodes reduces further job execution time of realistic documents (bigger documents), as compared with that of synthetic documents (smaller documents) processed with pipeline model. So, it is more advantageous to process bigger XML document using the pipeline model with more nodes. Regarding the number of CPU cores, 4 core processing is better than 1 core processing for both small (synthetic) and large (realistic) document processing. In Table 2, stock_doc is the smallest document in the three types of realistic document, and it has more XML tags than other documents. kernel_doc has more tags than scala_doc but most of them are empty tags. The empty tag processing needs less time than normal tag sets because the nodes can determine results of processing to empty tags earlier. Job execution time is sensitive to document types (number of tags and types of tags), node topology and processing type. Such characteristics also appears in other system indicators.

Regarding system active time (Figure 7 and 8), parallel processing is better than pipeline processing in all experiments. Regarding number of CPU cores, system active time is better for 4 core processing than for 1 core processing in both synthetic documents and realistic documents, regardless of document partition, processing allocation, node topology and buffer contention. The system active time also depends on the amount of processing to be executed.

Processing time is similar in both parallel processing and pipeline processing (Figures 9 and 10). Validation checking needs more time than well-formedness checking. The extra activity time in the pipeline processing is due to extra sending/receiving thread times. Regarding number of CPU cores, 1 core processing is better than 4 core processing in well-formedness checking (small processing). In contrast, 4 core processing is better than 1 core processing in validation checking (large processing) for synthetic documents (small document). But for some realistic documents (bigger document), 1 core processing has better validation checking processing time. If the number of tags that each node should process is roughly the same, 4 core processing leads to more buffer contention than 1 core processing. Generally, the system processing time reduces as the number of RelayNodes increases. However, sometimes (e.g. synthetic doc06) few RelayNodes are more efficient, due to specific document partition and processing allocation. Average processing time is greatly affected by whether we can allocate XML data efficiently. In addition, system processing time is also sensitive to

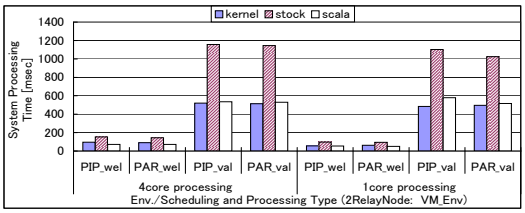


Figure 9: Sys. Processing Time (Real. Doc. Proc. by 2RN).

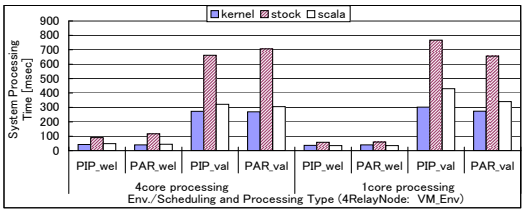


Figure 10: Sys. Processing Time (Real. Doc. Proc. by 4RN).

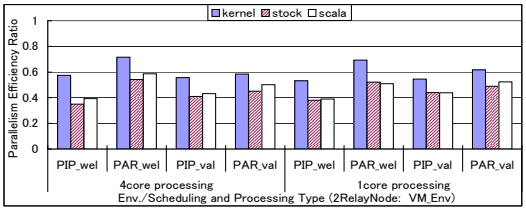


Figure 11: Parallel. Efficien. Ratio (Real. Doc. Proc. by 2RN).

XML document structure, number of tags and depth, which affects the amount of processing at each node.

Regarding parallelism efficiency ratio (Figures 11 and 12), 4 core processing is more efficient than 1 core processing excluding few cases of 2 RelayNode validation checking of realistic documents. In these cases, 1 core processing is a little more efficient.

Regarding system buffer access time (Figure 13 and 14), pipeline processing incurs larger buffer access time than parallel processing. In addition, most of 4 RelayNode processing incurs larger buffer access time than 2 RelayNode processing, because the amount of data passing through a node and assigned for process to the node is larger in the former case. In general, 1 core processing shows lower buffer contention than 4 core processing, as System buffer access time increases with more cores. Comparing the system buffer access time with other indicators, some indicators (e.g. system job execution time) are better for more CPU core processing but they also include inefficient performance because of useless waiting time caused by buffer contention.

Figures 15 and 16 further show node active time when processing doc01, for each node in the system.Generally, parallel processing is better than pipeline processing regarding node activity. This

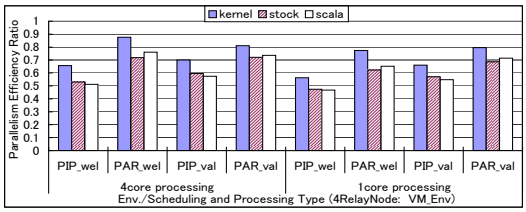


Figure 12: Parallel. Efficien. Ratio (Real. Doc. Proc. by 4RN).

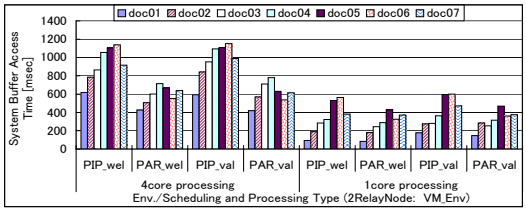


Figure 13: Sys. Buf. Access Time (Syn. Doc. Proc. by 2RN).

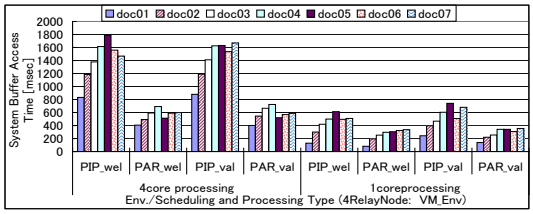


Figure 14: Sys. Buf. Access Time (Syn. Doc. Proc. by 4RN).

is because extra data transfer at each RelayNode is reduced in parallel processing, as compared with pipeline processing. Regarding the number of CPU cores in these figures, 4 core processing is better for StartNode, RelayNode and EndNode processing synthetic document. For MergeNode, 1 core performs better when processing synthetic documents. Across all node type, realistic document processing is superior when 4 cores are used.

Figures 17 and 18 further show node thread working time for processing kernel_doc. Generally, node thread working time is better for parallel processing using more RelayNodes. Regarding varying number of CPU cores, 4 core is better for StartNode when processing synthetic documents. However, 1 core RelayNode, EndNode and MergeNode present better node thread working time than their 4 core counterparts. Roughly, 1 core synthetic document processing is better in the following cases: less processing (e.g. well-formedness checking) and good document partition on RelayNodes; bad document partition, which leads to more processing for EndNode, for EndNode; Merge node, which leads to low buffer contention.

For convenience, we organize our performance characterization results about number of CPU cores

Table 3: Distributed XML Processing Characterization Summary.

	Synthetic docs (smaller docs)	Realistic docs (larger docs)
Job execution time	4 core is better	1 core is better in large proc.
System active time		4 core is better
System proc. time	1 core is better in WEL	Sometimes, 1 core is better in large proc.
Parallelism efficiency ratio	4 core is more efficient	4 core is more efficient in most cases
Sys. buffer access time	1 core is better	
Node active time	4 core is better in SN, RN and EN	4 core is better
Node thread working time	4 core is better in SN; Sometimes, 1 core is better in RN, EN and MN	4 core is better in SN; 1 core is better in EN; Sometimes, 1 core is better in RN and MN

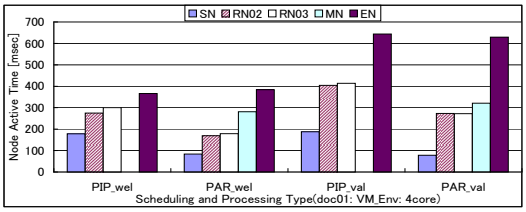


Figure 15: Node Ac. Time (doc01; 2RN; 4core).

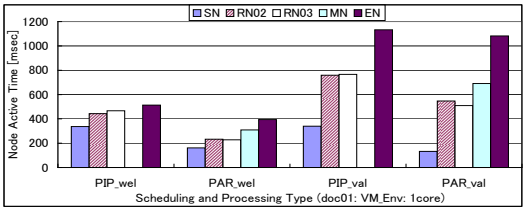


Figure 16: Node Ac. Time (doc01; 2RN; 1core).

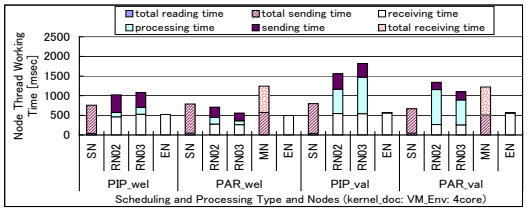


Figure 17: Node Th. Work. Time(kernel_doc; 2RN; 4core).

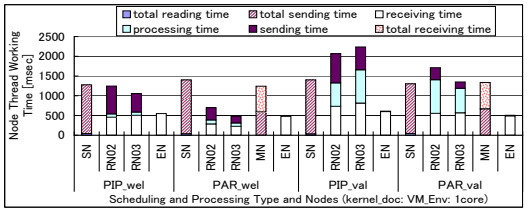


Figure 18: Node Th. Work. Time(kernel_doc; 2RN; 1core).

into Table 3 (WEL means well-formedness checking). A detailed analysis is also carried on at (Uratani et al., 2012), our previous work.

4 RELATED WORK

In this paper, we focus on offloading approach which assign a part of server's load to intermediate nodes in network for Web service. We here show some related works similar with our approach which have processing function at intermediate nodes. Active Network (Tennenhouse and Wetherall, 2007) is also focus processing at intermediate switches. The processing manner is described to two types: type-1) be capsuled into a packet, type-2) be assigned to the switches beforehand. The type-1 manner executes only simple processing but can process faster because of hardware execution. The type-2 manner, which like our system design, can describe even complicate processing. VNode(Kanada et al., 2012) also provides a processing environment at intermediate switches. In this research, the processing function is also provided at customized switches and its processing environment is provided as virtualized environment likewise using a virtual machine. These researches provide not only transport function but also processing function in the network. We address these intermediate processing model as a platform to achieve our distributed processing system.

Next we show current researches of processing in networks. In transcoding (Kim et al., 2012),a content server deliver data (e.g. video data) to clients via a transcoding server. The transcording server fixes the original data to another data which reflects user's demands. For instance, the data may be transformed from high resolution to low resolution at the transcoding server to adapt a mobile devices. A cache server(Nishimura et al., 2012; Kalarani and Uma, 2013) is a key technology of content delivery network. The cache servers are allocated to wide distributed places and they stock contents as cache from other content servers beforehand or with users request. Then the users' content request will lead to the nearest cache servers then the users receive contents with lower network latency. (Fan and Chen, 2012; Solis and Obraczka, 2006) focus sensor network. The researches propose to consolidate the large amount of sensing data at some intermediate nodes before the large data reach data collection servers.

Such approach can reduce energy consumption and network load for mobile sensor devices. (Shimamura et al., 2010) compress packets near a sender then extract the packets near a receiver during buffer queueing time to achieve low load network. In these researches and technology, we can assume the network provides special function such as video transforming function, data caching function and so on for a certain services.

5 CONCLUSIONS

In this paper, we have studied the impact of number of CPU cores in distributed XML processing using two models of distributed XML document processing in virtual environments for two types of XML document: parallel and pipeline models, on virtual machines with multicore CPUs, for synthetic and realistic XML documents. Regarding number of CPU cores for distributed XML processing, few CPU cores lead to less buffer contention. In contrast, more CPU cores leads to higher performance of some indicators, but with the drawback of incurring in wasteful buffer access waiting time. In addition, appropriate number of CPU cores depends on document characteristics. We can enhance the processing efficiency by improving buffer usage mechanism. As we have shown, pipeline processing is inefficient than parallel processing regardless document types and processing environment. The pipeline processing should treat parts of the document that are not to be processed at them. Such a specific node needs to be received and relayed to other nodes, consuming node resources and increasing processing overhead.

So far, we have focused on distributed well-formedness and validation of XML documents. These functions are a must for XML applications. The PASS-Node system guarantees the soundness of the XML document and it should lead to less battery consumption of mobile devices because of offloading. Moreover, other XML processing, such as filtering and XML transformations, can be studied. Internet routers in the future can do XML processing the same way routers today do deep packet inspection (Liu and Wu, 2013), as well as fast (hardware based packet) routing/forwarding.

We intend to study processing of streaming data other than XML documents at relay nodes (Shimamura et al., 2010). In such scenario, many web servers, mobile devices, network appliances, are connected with each other via an intelligent network, which executes streaming data processing on behalf of connected devices. The type of node processing is different than XML processing, given the less structured nature of streaming data, as compared with XML data.

ACKNOWLEDGEMENTS

Part of this study was supported by a Grant-in-Aid for Scientific Research (KAKENHI:24500043).

REFERENCES

Cavendish, D. and Candan, K. S. (2008). Distributed XML Processing: Theory and Applications. *Journal of Parallel and Distributed Computing*, 68(8):1054–1069.

Fan, Y.-C. and Chen, A. (2012). Energy Efficient Schemes for Accuracy-Guaranteed Sensor Data Aggregation Using Scalable Counting. *IEEE Transactions on Knowledge and Data Engineering*, 24(8):1463–1477.

Kalarani, S. and Uma, G. (2013). Improving the Efficiency of Retrieved Result through Transparent Proxy Cache Server. In *Proc. of fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) 2013*, pages 1–8.

Kanada, Y., Shiraishi, K., and Nakao, A. (2012). Network-virtualization Nodes that Support Mutually Independent Development and Evolution of Node Components. In *Proc. of IEEE International Conference on Communication Systems (ICCS) 2012, pages 363–367.*

Kim, S. H., Kim, K., Lee, C., and Ro, W. (2012). Offloading of Media Transcoding for High-quality Multimedia Services. *Consumer Electronics, IEEE Transactions on*, 58(2):691–699.

Liu, C. and Wu, J. (2013). Fast Deep Packet Inspection with a Dual Finite Automata. *IEEE Transactions on Computers*, 62(2):310–321.

Nishimura, S., Shimamura, M., Koga, H., and Ikenaga, T. (2012). Transparent Caching Scheme on Advanced Relay Nodes for Streaming Services. In *Proc. of International Conference on Information Networking (ICOIN), 2012*, pages 404–409.

Shimamura, M., Ikenaga, T., and Tsuru, M. (2010). Advanced Relay Nodes for Adaptive Network Services - Concept and Prototype Experiment. In *Proc. of International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA) 2010*, pages 701–707, Los Alamitos, CA, USA.

Solis, I. and Obraczka, K. (2006). In-network Aggregation Trade-offs for Data Collection in Wireless Sensor Networks. *Int. J. Sen. Netw.*, 1(3/4):200–212.

Tennenhouse, D. L. and Wetherall, D. J. (2007). Towards an Active Network Architecture. *SIGCOMM Comput. Commun. Rev.*, 37(5):81–94.

Uratani, Y., Koide, H., Cavendish, D., and Oie, Y. (2012). Distributed XML Processing over Various Topologies: Characterizing XML Document Processing Efficiency. In *Web Information Systems and Technologies*, volume 101 of *Lecture Notes in Business Information Processing*, pages 57–71. Springer Berlin Heidelberg.

Are Model-driven Techniques Used as a Means to Migrate SOA Applications to Cloud Computing?

Miguel Botto, Javier González-Huerta and Emilio Insfran
ISSI Group, Department of Information System and Computation, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, Spain
mibot@posgrado.upv.es, {jagonzalez, einsfran}@dsic.upv.es

Keywords: SOA, Cloud Computing, Migration.

Abstract: In recent years, cloud computing has emerged as an option to use computing resources as a solution for the global economic crisis, namely, a cheaper way to have IT resources. Thus, many companies have started to migrate their systems to cloud infrastructures, without the required support to carry out this process. In this position paper, we provide an overview of the current state of research on cloud computing migration. To understand this subject, we conducted a systematic mapping. The results suggest that research into cloud computing migration is still in its early stages. We identify research gaps and provide general recommendations about how these gaps may be addressed as well as future research directions that may have potential impact on this research field.

1 INTRODUCTION

Cloud computing has become a topic of tremendous interest as organizations struggle to improve their IT performance. Cloud computing is described by (Feuerlicht, 2010) as follows: it involves making computing, data storage, and software services available via the Internet.

The main characteristics of cloud services are: on-demand self-service, ubiquitous network access, location independent resource pooling, rapid elasticity, and measured service (Mell and Grance, 2011). Cloud computing is classified into service models as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). And their deployment models as public, private, community and Hybrid (Mell and Grance, 2011).

The use of cloud services enables companies to pay only for what they use with regard computing and network resources, rather than having to invest in IT resources, and the requisite staff to support all the hardware and software.

Cloud computing has associated benefits and also challenges. One of these challenges are related to is adoption, more specifically, the migration of existing application to cloud computing.

There are few studies as reported in (Mohagheghi and Sæther, 2011; Pace et al., 2010) that present the evaluation of different cloud platforms for performance indicators. Nevertheless, there is not sufficient literature available to support on process for migrating existing applications to cloud.

In this position paper, we present a systematic mapping study for summarizing how researchers and practitioners migrate their applications to cloud environments.

Our paper is organized as follows: Section 2, presents the protocol we defined. Section 3, described the results obtained. Finally, section 4 presents our conclusions and suggest areas for further investigation.

2 RESEARCH METHOD

We have performed a systematic mapping study by considering the guidelines that are provided in works as those (Budgen et al., 2007; Kitchenham and Charters, 2007; Petersen et al., 2008). A systematic mapping study is a means of categorizing and summarizing the existing information about a research question in an unbiased manner.

Our systematic mapping study was performed in three stages: Planning, Conducting, and Reporting. The activities concerning the planning and conducting stages of our systematic mapping study

are described in the following sub-section and the reporting stage is presented in Section 4.

Planning Stage. During the planning stage, we performed the following activities in order to establish a review protocol: establishment of the research question; definition of the search strategy, selection of primary studies, quality assessment, definition of the data extraction strategy, and selection of synthesis methods.

Research Question. The goal of our study is to examine the current use of strategies of migration of SOA applications to Cloud Computing environments from the point of view of the following research question: *"How researchers and practitioners migrate their SOA applications to Cloud Computing environments and which is the effect on the quality?"*. Since our research question is too broad, it has been decomposed into more detailed sub-questions (see Table 1) in order for it to be addressed.

Table 1: Research sub-questions.

Research sub-questions	Motivation
RQ1: Which strategies are used to migrate Service-Oriented Architecture applications to Cloud computing environments?	To discover which are the most frequently used migration strategies in the process to migrate SOA applications to Cloud Computing environments.
RQ2: Which are the consequences of migration on product quality?	To discover which quality characteristics have been affected in the migration process.

Search Strategy. The main digital libraries that were used to search for primary studies were: IEEEXplore, ACM Digital Library, Science Direct, and Springer Link. We also manually searched the following conference proceedings: CLOUD COMPUTING and IEEE CLOUD.

In order to perform the automatic search of the selected digital libraries, we used a search string (see Table 2) consisting of three parts: Migration, Services and Cloud. The period reviewed included studies published from 2006 to 2013 (inclusive). This starting date was selected because 2006 was the year in which Amazon Inc. officially launched *Amazon Web Services*.

Selection of Primary Studies. Each study was evaluated by the three authors in order to decide whether it should be included or not, by considering its title, abstract and keywords. The studies that met at least one of the following inclusion criteria were included:

- Research papers presenting migration strategies SOA applications to Cloud Computing environments.
- Research papers presenting examples or any empirical studies (e.g. study cases, experiments), about migration strategies to Cloud Computing environments.

Table 2: Search string applied.

Concept	Alternative terms or Synonyms
Migration	(migra* OR evolv* OR adopt* OR reus* OR mov*) AND
Services	(soa OR service*) AND
Cloud	cloud

* The asterisk symbol "*" signifies any character whose purpose it is to include any word variation of each search term (e.g., the search term "migra*" includes the following words: migrate OR migrating OR migration OR…).

The studies that met at least one of the following exclusion criteria were excluded:

- Introductory papers for special issues, books and workshops.
- Duplicate reports of the same study in different sources.
- Short papers with less than five pages.
- Paper not written in English.

Quality Assessment. In addition to general inclusion/exclusion criteria, it is considered critical to assess the "quality" of primary studies. A three-point Likert-scale questionnaire was designed to provide a quality assessment of the selected studies. The questions were:

- The study presents strategies to migrate SOA applications to Cloud Computing environments.
- The study has been published in a relevant journal or conference.
- The study has been cited by other authors.

The possible answers to these questions were: "I agree (+1)", "Partially agree (0)", and "I do not agree (-1)".

Data Extraction Strategy. It was based on providing the set of possible answers for each research sub-question that had been defined. The possible answers to each research sub-question are explained in more detail as follows.

With regard to RQ1 (Strategies used), a paper can be classified in one of the following answers:

- C1: Migration strategies.
  a) *Conventional*: if it paper uses existing conventional migration strategy.
  b) *MDD*: if it paper uses strategies based on Model-Driven Development approach.
- C2: Migration approaches (Watson, 2012).

a) *Rehost:* if migration approach is specifically to move an application without changing its architecture.
b) *Refactor:* if migration approach is specifically to move applications to a different hardware environment and/or change the application infrastructure configuration without changing its external behavior.
c) *Revise:* if migration approach is specifically to modify or extend the existing code base to support legacy modernization requirements.
d) *Rebuild:* if migration approach is specifically to rebuild the solution, discard code for an existing application and re-architect the application.

- C3: Migration types (Andrikopoulos et al., 2012).
  a) *Replace components:* a migration type where one or more (architectural) components are replaced by Cloud services.
  b) *Partially migrate:* if migration type is specifically to change some of the application functionality to the Cloud, such as application layers, and architectural components.
  c) *Migrate the whole software stack:* if migration type is specifically to move the application that is encapsulated in VMs and run them on the Cloud.
  d) *Cloudify:* if migration type is specifically to make a complete migration of the application tasks.

With regard to RQ2 (C4: Quality aspects), a paper can be classified in one or more quality characteristics from the ISO/IEC 25010 standard SQuaRE (International Organization for Standardization, 2011).

Synthesis Method. We applied both quantitative and qualitative synthesis methods. The quantitative synthesis was based on: i) Counting the primary studies classified in each answer from our research sub-questions; and ii) Counting the number of papers found in each bibliographic source per year.

The qualitative synthesis is based on including several representative studies for each criterion by considering the results from the quality assessment.

Conducting Stage. The application of the review protocol yielded the following preliminary results (see Table 3): A total of 48 research papers were therefore selected in accordance with the inclusion criteria.

Table 3: Results of conducting stage.

Source	Potential Studies	Selected Studies
Automated search		
IEEEXplore (IEEE)	1026	29
ACM DL (ACM)	308	7
Science Direct (SD)	152	5
Springer Link (SL)	1187	2
Total	2673	43
Manual search		
CLOUD COMPUTING	8	1
IEEE CLOUD	7	4
Total	15	5
Overall results	2688	48

3 RESULTS

A summary of the results of our study is presented in Table 4. Table 5 includes the papers cited in this section. The complete list is available at: http://users.dsic.upv.es/~jagonzalez/Webist2014/references.htm

Table 4: Results of systematic mapping.

Criteria	Possible answer	Results	
		# Studies	Percentage (%)
C1: Migration strategies	Conventional	46	95,83
	MDD	2	4,17
C2: Migration approach	Rehost	20	44,44
	Refactor	18	40,00
	Revise	-	-
	Rebuild	7	15,56
C3: Migration types	Replace	-	-
	Partially migrate	7	14,89
	Migrate the whole software stack	15	31,91
	Cloudify	25	53,19
C4: Quality aspects	Performance efficiency	44	30,56
	Compatibility	14	9,72
	Reliability	31	21,53
	Security	35	24,31
	Maintainability	5	3,47
	Portability	15	10,42

3.1 Migration Strategies

With regard to criteria C1 (migration strategies) revealed that around 96% of the papers reviewed presented conventional strategy (e.g., Babar et al. on Table 5.S01, and Tran et al. on Table 5.S14).

Babar et al. (see Table 5.S01) reported their

Table 5: Excerpt of selected papers from the review.

S01. Babar MA, Chauhan MA. A tale of migration to cloud computing for sharing experiences and observations, SECLOUD, 2011.
S02. Beserra P V., Camara A, Ximenes R, Albuquerque AB, Mendonca NC. Cloudstep: A step-by-step decision process to support legacy application migration to the cloud, MESOCA, 2012.
S03. Cai B, Xu F, Ye F, Zhou W. Research and application of migrating legacy systems to the private cloud platform with cloudstack. ICAL, 2012.
S04. Chauhan MA, Babar MA. Towards Process Support for Migrating Applications to Cloud Computing, International Conference on Cloud and Service Computing, 2012.
S05. Chee Y-M, Zhou N, Meng FJ, Bagheri S, Zhong P. A Pattern-Based Approach to Cloud Transformation, International Conference on Cloud Computing, 2011.
S06. Gerhards M, Sander V, Belloum A. About the flexible Migration of Workflow Tasks to Clouds Combining on- and off-premise Executions of Applications, Cloud Computing, 2012.
S07. Guillen J, Miranda J, Murillo JM, Canal C. A service-oriented framework for developing cross cloud migratable software. NordiCloud, 2013.
S08. Guillen J, Miranda J, Murillo JM, Canal C. Developing migratable multicloud applications based on MDE and adaptation techniques, NordiCloud, 2013.
S09. Juan-Verdejo A, Baars H. Decision support for partially moving applications to the cloud, HotTopiCS, 2013.
S10. Lamberti F, Sanna A, Demartini C. How to move your own applications into the cloud by exploiting interfaces automation and accessibility features, International Conference on Cloud Computing and Intelligence Systems, 2011.
S11. Li J, Jia Y, Liu L, Wo T. CyberLiveApp: A secure sharing and migration approach for live virtual desktop applications in a cloud environment. Futur. Gener. Comput. Syst. 2013.
S12. Mohagheghi P, Sæther T. Software Engineering Challenges for Migration to the Service Cloud Paradigm: Ongoing Work in the REMICS Project, IEEE World Congress on Services, 2011.
S13. Suen C-H, Kirchberg M, Lee BS. Efficient Migration of Virtual Machines between Public and Private Cloud, Inter. Conf. on Cloud Computing Technology and Science, 2011.
S14. Tran V, Keung J, Liu A, Fekete A. Application migration to cloud, SECLOUD, 2011.
S15. Zhou L. CloudFTP: A Case Study of Migrating Traditional Applications to the Cloud, Inter. Conf. on Intelligent System Design and Engineering Applications, 2013.

experiences and observations gained from migrating an Open Source Software, Hackystat, to cloud computing. The aims of this job is to share their experiences and observations gained through this project and to analyze the literature with those research works who intend to migrate software systems in general and SOA based system in particular to cloud computing.

Tran et al. (see Table 5.S14) proposed a taxonomy of the migration tasks involved, and they showed the costs breakdown among categories of tasks, for a case-study which migrated a .NET n-tier application to Windows Azure. They determined how efforts are required for different types of Cloud as well as identified any missing tasks.

The remaining 4% of the studies reported the use of *MDD* strategy. MDD approaches rely on models as a means of abstracting the development process from the peculiarities of each cloud platform. These results may indicate that there are few studies that used this strategy to migrate existing system to cloud computing environment (e.g., Guillen et al. on Table 5.S08, and Mohagheghi et al. on Table 5.S12).

Guillen et al. (see Table 5.S08) proposed a framework MULTICLAPP. The framework follows a three-stage development process where applications can be modeled and coded without developers having to be familiar with the specification of any cloud platform.

Mohagheghi et al. (see Table 5.S12) presented a research project REMICS to define methodology and tools for model-driven migration of legacy applications to a SOA with deployment in the cloud. The project's main objective is to develop a set of model-driven methods and tools that support organizations with legacy systems to modernize them according to the "Service Cloud paradigm".

3.2 Migration Approaches

With regard to criteria C2 (migration approach) revealed that the most frequently used migration approach is *rehost*, with around 44% of the papers reviewed (e.g., Li et al. on Table 5.S11 and Zhou et al. on Table 5.S15). These results may indicate that most migrations are performed using this approach.

Li et al. (see Table 5.S11) proposed a flexible collaboration approach, CyberLiveApp, to enable live virtual desktop application sharing, based on a cloud infrastructure. This approach supports secure application sharing and on-demand migration among multiple users or equipment. To achieve the goals of live application sharing and migration between VMs, a presentation redirection approach based on VNC protocol and a VM cloning service based on the Libvirt interface are used.

Zhou et al. (see Table 5.S15) proposed migrating traditional applications - CloudFTP to the cloud. They implemented FTP service on Windows Azure Platform along with the auto-scaling cloud feature since CloudFTP follows the application model suggested for general Azure development.

Refactor account for around 40% of the papers reviewed (e.g, Beserra et al. on Table 5.S02, and Chee et al. on Table 5.S05).

Beserra et al. (see Table 5.S02) presented Cloudstep, a step-by-step decision process aimed at supporting legacy application migration to the cloud. The process relied on the creation of template-based profiles characterizing the organization, the target legacy application and candidate cloud providers.

Chee et al. (see Table 5.S05) proposed a pattern-based approach Cloud Transformation Advisor which helps users selecting appropriate enablement patterns from a knowledge base of best practices when planning the transformation. This knowledge base uses a structured representation to capture application information, cloud platform capability information, and enablement pattern information in order to facilitate pattern selection.

Rebuild account for around 16% of the papers reviewed, since this way focused on rebuilding the solution (e.g., Cai et al. on Table 5.S03).

Cai et al. (see Table 5.S03) presented the problems with Legacy Information Systems and proposed redevelopment, which rewrites existing applications providing new interfaces for a component, making it more easily accessible for other software components; and migration.

3.3 Migration Types

With regard to for criteria C3 (migration types) revealed that around 53% of the papers reviewed presented *cloudify* (e.g., Chauhan et al. on Table 5.S04, and Lamberti et al. on Table 5.S10).

Chauhan et al. (see Table 5.S04) presented a process framework for supporting migration to cloud computing based on their experiences from migrating an Open Source System (OSS), Hackystat, to two different cloud computing platforms (Amazon Web Services and Google App Engine). The main activities involved in this process include identification of requirements and potential cloud platforms, analyzing application compatibility with potential cloud environments, identification of potential architecture solutions, evaluation of cloud environments for cloud specific quality attributes, tradeoff analysis of potential architecture solutions, selection of architecture modifications to be incorporated and refactoring of the system to incorporate new architecture modifications.

Lamberti et al. (see Table 5.S10) presented a prototype implementation, which consists of a framework that allows today's applications to be put into the cloud by exploiting user interface automation and accessibility information embedded in modern window-based graphics toolkits. This framework organized as a three-tier architecture includes the particular remote program we want to move to the cloud, a server-side gateway and a web-based client.

Migrate the whole software stack account for around 32% of the papers reviewed (e.g., Suen et al. on Table 5.S13). Suen et al. (see Table 5.S13) proposed and evaluated techniques for both instance-based and volume-based storage in the public and private cloud infrastructure for efficient and effective transfer and storage of VM images. The main focus was on both the public and private cloud infrastructure and the movement of VMs between them.

Lastly, *partially migrate* account for around 15% of the papers reviewed (e.g., Gerhards et al. on Table 5.S06, and Juan-Verdejo et al. on Table 5.S09).

Gerhards et al. (see Table 5.S06) addressed the demand for a consistent framework that allows a mixture of on and off-premise calculations by migrating only specific parts to a Cloud. It used the concept of workflows to present how individual workflow tasks can be migrated to the Cloud whereas the remaining tasks are executed on-premise.

Juan-Verdejo et al. (see Table 5.S09) proposed a cloud migration framework to assist in the moving of the target application by following a local and cloud deployment model instead of an all-or-nothing approach. Besides this framework envisions that parts of the application are kept locally while others parts are migrated to cloud infrastructures.

3.4 Quality Aspects

With regard to criteria C4 (quality aspects) revealed that the most frequently quality aspects were *performance/efficiency* and *security* account for around 31% and 24% respectively (Guillen et al. (see Table 5.S07)) since on cloud environments the elasticity property of applications where quick and secure deployment is typically required. Others quality aspects as *maintainability* and *compatibility* account for around 3% and 9% respectively. This is because of quality aspects such as maintainability to play a minor role because the cloud providers are responsible of this partly on their platforms.

On the other hand, *reliability* and *portability* account with 21% and 10% respectively received less considerations (e.g., Babar et al. (see Table 5.S01)), where the authors stated that one example of such requirement is constrains on the geographic locations of data storage places. For some applications, it is required that data should not store outside a particular region. In IaaS clouds, data is

maintained at different location for reliability and cost effectiveness.

It is worthy to mention that the analysis of the number of research studies on cloud migration showed that there has been a growth of interest on this topic since 2009. Figure 1 shows the number of selected publications per year and source.

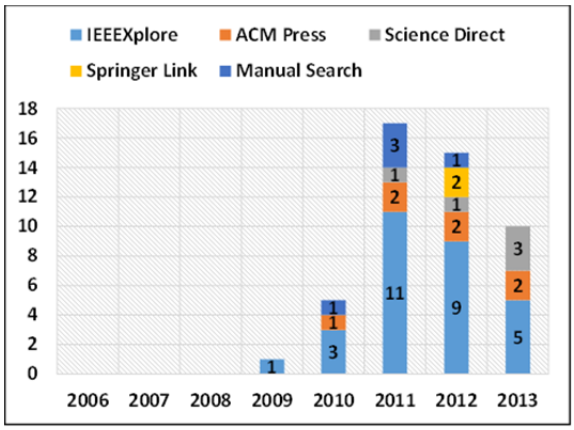


Figure 1: Number of publications by year and source.

4 CONCLUSIONS

The results suggest that research into cloud computing migration is still in its early stages. We identify research gaps and provide general recommendations about how these gaps may be addressed as well as future research directions that may have potential impact on the field.

Some of the gaps found are: i) MDD approach had been rarely used in the process to migrate SOA applications to Cloud environments; and ii) Some quality characteristics which we consider relevant in applications (reliability, maintainability, portability) had not received appropriate coverage.

Our results also confirmed some claims stated by other researchers according to the cloud migration: there is a lack of literature to support the migration process. We recommend the use of Model-Driven Development approach since it provides a higher abstraction level than traditional programming languages. We are also intended to address some of the research gaps discovered. For instance, addressing migrate SOA applications to Cloud environment through a migration framework using MDD techniques taking advantage of its benefits.

ACKNOWLEDGEMENTS

This research was supported by the Senescyt-Ecuador (scholarship program 2011) and ValI+D fellowship program (ACIF/2011/235).

REFERENCES

Andrikopoulos, V.; Binz, T.; Leymann, F.; Strauch, S., 2012. How to adapt applications for the Cloud environment. *Computing*, 95(6), pp.493–535.

Budgen, D.; Turner, M.; Brereton, P.; Kitchenham, B., 2008. Using Mapping Studies in Software Engineering. In *Proceedings of PPIG 2008*. Lancaster University, pp. 195–204.

Feuerlicht, G., 2010. Next Generation SOA: Can SOA Survive Cloud Computing? In *Advances in Intelligent Web Mastering-2*. pp. 19–29.

International Organization for Standardization, 2011. ISO/IEC 25010: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.

Kitchenham, B. & Charters, S., 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering*,

Mell, P. & Grance, T., 2011. The NIST Definition of Cloud Computing, Recommendations of the National Institute of Standards and Technolog. *National Institute of Standards and Technology*.

Mohagheghi, P. & Sæther, T., 2011. Software Engineering Challenges for Migration to the Service Cloud Paradigm: Ongoing Work in the REMICS Project. In *2011 IEEE World Congress on Services*. IEEE, pp. 507–514.

Pace, E.; Betts, D.; Densmore, S; Dunn, R., 2010. *Moving Applications to the Cloud on the Microsoft Azure Platform*, Microsoft Press.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M., 2008. Systematic mapping studies in software engineering. *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pp.68–77.

Watson, R., 2012. How to migrate applications to the cloud. *Information Age*. Available at: http://www.information-age.com/technology/applications-and-development/1681988/how-to-migrate-applications-to-the-cloud [Accessed October 1, 2013].

Designing a General Architecture for Data Interchange

Alina Andreica[1], Josef Küng[2], Gabriela Şerban Czibulla[1] and Christian Sacarea[1]
[1]*Babeş-Bolyai University, Cluj-Napoca, Romania*
[2]*Johannes Kepler Uniersity, Linz, Austria*
alina.andreica@ubbcluj.ro, jkueng@faw.jku.at, gabis@cs.ubbcluj.ro, csacarea@math.ubbcluj.ro

Keywords: Data Interchange Model, Academic Data Exchange, Software Solution.

Abstract: The paper describes principles for designing a general framework for automatic data interchange that scopes all three levels, data, semantic and knowledge. In spite of the huge amount of research already performed and existing standards and products, there is room to enhance information and knowledge integration. Consequent to defining the data interchange framework, we are going to apply these principles in developing and implementing a solution for academic data interchange. Such a solution has the potentiality for important advantages in academic cooperation and societal benefits.

1 INTRODUCTION AND WORKING FRAMEWORK

Interoperability is the capability of different systems to share functionalities or data (Olmedilla et al, 2006). System interoperability has been dealt with by means of various models (Morris et al, 2004) and has been extensively researched for business processes (Ziemann, 2010).

Interoperability may be achieved by: *system integration* (Chapman, Kihn, 2009), which has also been mainly tackled in the literature for business and organizational processes (Hasselbring, 2000), (Hasselbring, Pedersen, 2005), a *semantic approach*, or *data exchanges*, with important results for business processes. Interoperability issues have to be considered on a three layer basis.

1. A first layer concerns *data interoperability*. For this, several standards have been developed, like XML and SQL standards. They also solve syntactic interoperability issues. In the last decade, many instruments have been developed, mainly driven by the necessity of solving Business-to-Business and e-commerce problems. For instance, ebXML (OASIS, 2007) has been developed as a new, global standard for Internet-based B2B e-commerce (Gibb, Damodaran, 2002). Also, several specifications have been developed, like AS2 for secure and reliable transport of data over the Internet, AS4 (AS4 web, 2012) for B2B documents exchange using Web services, using XML encryption and XML Digital Signature, or ebMS, a communication mechanism which has to be implemented for business document exchange within the ebXML standard.

The Electronic Data Interchange (EDI) model (Adams et al, 2002) was proposed for standardizing business information exchange, various formats being used in this respect: ANSI X.12, XML (cXML, xCBL, OpenTrans, UBL).

XML (eXtended Markup Language) (XML web, 2010) is a widely used standard for information structuring and information exchange. XML is often used as a format for data exchange and integration between web applications or services within organizations, and accordingly, integration of XML data has become an important research problem. Studies regarding the XML definitions dedicated to representing database structures already lead to a XML representation of a database and algorithms for obtaining it (Abiteboul et al, 2000), (Bourret, 2000).

Analytical processing of data distributed in the World Wide Web is an important topic for current research (Heflin, Stuckenschmidt 2012). XML data integration (Le et al, 2006), (Algergawy et al, 2010) or RDF-based integration (Karnstedt et al, 2012) involves reconciliation at different levels: (1) at schema level, reconciling different representations of the same entity or property, and (2) at instance level, determining if different objects coming from different sources represent the same real-world entity. An important problem to be solved for XML-based data integration is the support for integrated access to the data using advanced query processing capabilities (Manolescu et al., 2001), (Sattler et al,

2005). The unique correspondence between the database schema and its XML representation is tackled and proved in (Andreica et. al, 2005). Current approaches for analytical processing on integrated (Algergawy et al, 2011), (Heise, Naumann, 2011) or distributed (Vouros et al, 2010) data, sometimes based on cloud or grid architectures, give good frameworks for application development.

The learning objects model LOM and the SCORM standard (SCORM web, 2012) have been improved in various directions: personalized adaptive learning frameworks based on user profiles (Arroyo et al, 2006); *activating learning objects* with DBLink (Kassahun et al, 2006); describing how learning objects should be used within VC-LOM (Virtual Campus-LOM) (Di Nitto et al, 2006); universal interoperability layer for educational networks with Simple Query Interface (SQI).

Currently, various learning standards (Walker, 2012) are being used: standards for moving content (IMS Common Cartridge (CC-MS web, 2012), SCORM 2004's Content Management Component), data standards, like SIF Association www.sifinfo.org - and PESC (PESC-CEDS web, 2012), integration standards - PESC's Data Transport Standard, SIF (SIF Specifications web, 2012), IMS Learning Tool Interoperability (LTI-IMS web, 2012), AICC-CMI (AICC web, 2012) and some of them creating interoperability contexts, such as SCORM, DCMI, PESC or Ed-Fi. The Systems Interoperability Framework (SIF) Association has achieved relevant results, proposing specifications for event reporting, data provisioning, messages and agents.

Among messaging standards, we enumerate: RosettaNet (Rosetta web, 2011), ebMS (ebXML Messaging Service) (OASIS, 2007), AS2 (Applicability Statement 2) (AS2 web, 2012).

2. The second layer is the *semantic layer*. Semantic interoperability is the ability to share the meaning of electronic documents using computer systems. Knowledge discovery, inference, logic are enabled by semantic interoperability. Semantic interoperability is mainly achieved using ontologies (Dicheva et al, 2005), (Yuan et al, 2010).

The Resource Description Framework (RDF) (Lassila, Swick, 1999), a data model of metadata instances, is recommended by the W3C (World Wide Web Consortium) in order to solve semantic interoperability problems. SHOE (Simple HTML Ontology Extensions) (Heflin, Hendler, 2000) introduces an ontology-based knowledge representation language designed for the Web that supports interoperability by sharing and reusing ontologies. The Open Group (Open web, 2012) focuses on developing a particular open standard to facilitate semantic interoperability: the Universal Data Element Framework – UDEF (UDEF web, 2012). UDEF framework is integrated with the Resource Description Framework (RDF) and especially used for describing business operations.

3. The third layer is the *knowledge layer* and it addresses the conceptual structure of the shared data, information and knowledge. Knowledge sharing over computer information systems is a major task for the interoperability approach. The first step is to identify and automatically discover knowledge. This process is based on a common cultural and social background, enabling the user to identify and use the extracted knowledge. Technically, the process uses the previous two layers, and is based on the Conceptual Knowledge Processing paradigm (Stumme, Wille 2000). Transparency about the identified knowledge has to be ensured. Active knowledge systems enable to capture and represent knowledge, as well as to reason and to draw appropriate conclusions (Hitzler, Schärfe, 2009).

We consider that the Open Internet of Things standards (OpenIoT, 2013) may be used as an efficient framework for data interchange.

2 DATA INTERCHANGE PRINCIPLES

We aim at proposing a general data interchange model as a data interchange and interaction model, valid for various fields. The framework we design is based on the principles described below.

We use standards for the agent based system development, and for the communication between the agents within the multi-agent system, in the form of FIPA-ACL (FIPA web) messages.

We adapt agent-oriented methodologies - like Gaia (Wooldridge et al, 2000) - for MAS development influence interoperability within the system. Ontologies (Gruber, 1993) are used for achieving semantic interoperability in the multi-agent educational system. Ontologies are powerful tools for sharing knowledge sources in a scalable, adaptable and extensible manner and as well for reaching semantic interoperability among heterogeneous, distributed systems.

We use a multi-agent architecture (Weiss, 1999) for designing the data exchange model proposal in order to benefit from the advantages that agent based technology offers: decentralization, extensibility, robustness, maintainability, flexibility. In order to

preserve confidentiality of the interchanged data, as well as for security reasons, the mobile agent (Nwana, 1996) technology is used. We provide self-adapting communicating objects, which work on distributed datasets not only supporting the exchange but also the analysis of distributed sources.

The knowledge layer is based on the Conceptual Knowledge Processing paradigm, which makes use of concept lattices, i.e., knowledge maps displaying concepts and their hierarchies, with a clear semantic and a very high expressivity. The methods used in this respect are knowledge discovery, knowledge acquisition, knowledge development, knowledge distribution and knowledge sharing. They are based on Formal Concept Analysis (Ganter et al, 2005), the mathematical theory of concepts and their hierarchies, which is a widely accepted standard of knowledge processing and representation.

We use the Open Internet of Things standards (OpenIoT, 2013) as a framework for data interchange; the "Utility/Application Plane" and "Virtualized Plane" (OpenIoT, 2013) layers provide a flexible framework for information communication and exchange, including cases of cloud hosted data.

3 ACADEMIC APPLICATION CASE

A specific application case of the data interchange model we are developing is the academic field. In this respect, *system inter-operability* capabilities, based on data interchange techniques, provide support for student or teacher mobility and e-learning content exchange, relevant for networks of universities with common programs. We note that e-learning content exchanges are important mainly in partner educational networks.

We take into account specific academic information, such as study levels, specializations, courses, students, teachers, grades, as well as e-learning content resources. This information standardisation has important benefits for exchanging information in various systems, including cloud services. Efficient data interchange components increase academic activity proficiency, with important societal benefits.

The proposed model for academic information exchange –fig. 1, supports semantic interoperability. This framework may also be used for advanced data analysis. We propose a Reference Interaction Model for defining the data content that is needed to provide an explicit representation of semantic and lexical connections that exist between academic entities and we implement intelligent agents. The interaction framework among the functional components of the academic information system environment is built using actors and transactions.

We define Integration Profiles for sharing information within academic institutions and across networks. Integration Profiles address data analysis and interoperability issues related to information access for academic actors and students, academic workflow, security and administration infrastructure, as well as for potential community or business actors. Each profile defines the actors, transactions and information content required to fulfil common interactions between academic entities or services provided to community stakeholders.

Multi-agent systems (MAS) are appropriate for modelling the academic domain, which involves interactions between various organizations with different (possibly opposite) goals (Shen, Barthes, 2001), where flexible autonomous actions are required for achieving the goals. Ontologies and agent technologies may be combined in order to successfully enable heterogeneous knowledge sharing. We develop components for exploring and reasoning on large-scale educational data to better understand learners' educational evolution, assess their progress and evaluate learning environments.

We use abstract methods for agents' database access (Han, Kamber, 2006), and particularize them for certain standard academic information system technologies. We exploit the advantages of using intelligent agents as a support for the active data mining (Agrawal, Psaila, 1995). For example, when new data is added, a triggering agent can notify the main mining application, so that new data can be compared to the already mined data. Another scenario regards sending alerts and notifications in critical situations (i.e., possible frauds).

Data mining (Han, Kamber, 2006) and knowledge discovery techniques are used to find interesting relationships and patterns within academic data (e.g. attributes of students, assessments). Machine learning techniques, such as reinforcement learning (Sutton, Barto, 1998), are useful for learning teaching strategies in an adaptive and intelligent educational system. Intelligent triggering software agents are efficient for sending alerts in critical situations (i.e., frauds, etc.).

We are implementing distributed data mining techniques (Domenico, Trunfio, 2010) (e.g. fraud detection, students' profiling) within the educational system. The agent based model is used within the distributed data mining architecture, the mobile

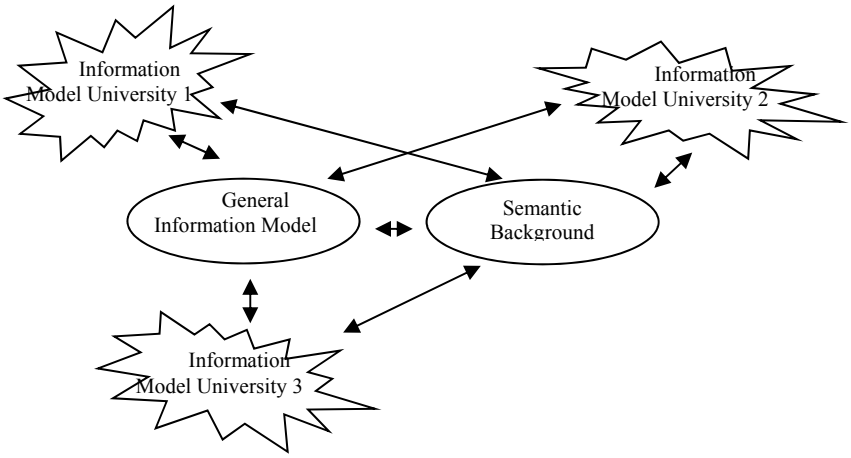


Figure 1: Heterogeneous sources, General Information Model and Semantic Background Knowledge.

agent model being useful for preserving academic data confidentiality.

Machine learning techniques (Mitchell, 1997) are used in order to enrich the educational system with the characteristics of intelligence and adaptability. We develop and use computational intelligence and machine learning techniques (e.g. learning and teaching, intelligent tutoring, adaptive information filtering, etc): *relational association rules mining* (Şerban et al, 2008) for predictive modelling (e.g predicting the most appropriate specialization for a student, predicting the students' grades, etc); *fuzzy* to deal with imprecision, uncertainty, partial truth; *unsupervised classification techniques*, such as clustering (Jain, 2010) and self-organizing maps (Kohonen et al, 2001) in order to uncover hidden patterns within academic data (students' profiling,identifying groups of students sharing common interests, etc); reinforcement learning techniques (Sutton, Barto, 1998) in order to discover users' (students, teachers, etc.) preferences, to recommend specific tasks to students, according to their preferences, to develop optimal teaching strategies by adapting tutoring to students' needs; *approaches that ensure privacy* - e.g. (Tran, Küng, Quoc 2011) proposes a particular k-anonymity technique that does not affect the association rule mining quality.

We are developing intelligent techniques for data cleaning within academic data collections, using techniques that detect and correct data errors. Data cleaning is an important pre-processing step in a data mining process; various computational intelligence techniques such as relational association rules (Şerban et al, 2008) are used in this respect.

Existing SIF (SIF Specifications web) standards are adapted for performing the exchange, management and integration of electronic academic information. It is important to develop means and specifications that ensure messaging standards for academic transactions in order to achieve interoperability. Such standards increase the effectiveness and efficiency of academic information delivery within and among academic organizations.

4 CONCLUSIONS

We describe the principles we are applying in designing a general framework for data interchange between information systems. The general data interchange model is extremely useful for ensuring entity cooperation in various fields. The model will be applied for the academic field.

The data interchange model we design is going to provide: *advanced data exchange services* using self-adaptive software communicating objects which provide academic IT services within distributed architectures; *a standardization framework* which supports information management and data exchange, ensuring interoperability in software technologies and services; *a* platform independent *software solution for academic data exchange*, with important social and collaborative advantages – as a flexible software tool for sustaining academic communication and cooperation. The solution meets all privacy requirements of academic institutions and national laws.

The application case for the academic field will provide important advantages for increasing academic competitiveness, with a significant societal impact on academic institutional cooperation, student and teacher exchanges, efficient information

management and access to data analysis facilities for academic and community stakeholders. A relevant advantage of the solution is its flexibility and efficiency, including real-time response features, efficient information exchange (only relevant data is exchanged), with minimal resources involved.

REFERENCES

Abiteboul, S., Hull, R., Vianu, V., 1995. *Foundations of Databases*, Addison-Wesley Publishing Company.

Abiteboul, S., Buneman, P., Suciu, D., 2000. *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann Publishers.

Adams, S.; Hardas, D.; Iossein, A. and Kaiman, C., 2002. *BizTalk Unleashed*. Indianapolis, Indiana: Sams Publishing. p. 966.

Agrawal, R., Psaila, G., 1995. *Active Data Mining*, KDD 1995, p. 3-8.

Aggarwal, R; Charu C.; Wang, H. (Eds.), 2010. *Managing and Mining Graph Data*, Advances in Database Systems 40, Springer.

AICC – CMI Guidelines for Interoperability AICC http://aicc.org/docs/tech/cmi001v4.pdf, Acc Dec 2012.

Algergawy, A., Mesiti, M., Nayak, R.; Saake, G., 2011. *XML data clustering: An overview*, ACM Comput. Surv. 43(4), p. 25.

Algergawy, A., Nayak, R; Saake, G, 2010. *Element similarity measures in XML schema matching*. Inf. Sci. 180(24), p. 4975-4998.

Andreica, A-B; Stuparu, D; Mantu, I**,** 2005. *Symbolic Modelling of Database Representations*, Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2005, P.59 - 62.

Aroyo, L., Dolog, P., Houben, G-J., Kravcik, M., Naeve, A., Nilsson, M., Wild, F., 2006, Interoperability in Personalized Adaptive Learning, *Educational Technology & Society*, 9 (2), 4-18.

AS2 –MIME-based Secure Peer-to-Peer Business Data Interchange Using HTTP, Applicability Statement 2 (AS2), RFC 4130, http://tools.ietf.org/html/rfc4130 , Accessed December 2013.

AS4 – AS4: Web Services for B2B, GS1 eTG White Paper, 2011 http://www.gs1.org/docs/ecom/AS4_-_A_new_tools_for_the_B2B_toolbox.pdf.

Bourret, R. P., 2000. *XML Database Products*, http://www.rpbourret.com/xml/ , acc November 2011.

CC-IMS – "Common Cartridges Frequently Asked Questions", *IMS Global Learning Consortium,* http://www.imsglobal.org/cc/ccfaqs.html#3, Dec 2012.

Chapman, C, Kihn, L-A., 2009. Information system integration, enabling control and performance, *Accounting, Organizations and Society*, Volume 34, Issue 2, P. 151–169.

Chen, J Y; Lonardi, S (Eds), 2009. *Biological Data Mining*, Chapman and Hall/CRC; 1st edition.

Dang, T. K., Küng, J., Van Quoc, P. H., 2011. *Protecting Privacy While Discovering and Maintaining Association Rules*,Conference: New Technologies, Mobility and Security - NTMS, IEEE, p 1-5.

Dicheva, D., Sosnovsky, S., Gavrilova, T., Brusilovsky, P., 2005. Ontological Web Portal for Educational Ontologies, *Proc. of the Workshop on Applications of Semantic Web Technologies for e-Learning.*

Di Nitto, E., Mainetti, L., Monga, M., Sbatella, L., Tedesco, R., 2006. Supporting Interoperability and Reusability of Learning Objects: The Virtual Campus Approach, *Educational Tech. & Society*, 9 (2), 33-50.

FIPA – The Foundation for Intelligent Physical Agents, www.fipa.org , Accessed December 2012.

Ganter, B., Stumme, G., Wille, R., 2005. *Formal Concept Analysis - Foundations and Applications*, Springer.

Gibb, B. K., Damodaran, S., 2002. *ebXML: Concepts and Applications*, John Wiley & Sons, Inc. New York.

Gruber, T. R. 1993, *A translation approach to portable ontology specifications*. Knowledge Acquisition, volume 5, pp. 199–220.

Han, J., Kamber, M., 2006. *Data Mining: Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers.

Han, J.; Gao, J., 2009. *Research Challenges for Data Mining in Science and Engineering*, H. Kargupta et al., (eds.), Next Generation of Data Mining, Chapman & Hall, 3-28.

Han, J., Kamber, M., Pei, J., 2011. *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann.

Hasselbring, W., 2000. "Information System Integration"*, Communications of the ACM*, 43 (6). pp. 32-36.

Hasselbring, W., Pedersen, S., 2005. "Metamodelling of Domain-Specific Standards for Semantic Interoperability", *Lecture Notes in Computer Science*, 3782, pp. 557 – 559.

Heflin, J., Hendler, J., 2000. *Semantic Interoperability on the Web* http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf.

Heflin, J., Stuckenschmidt, H., 2012, *Web-scale semantic information processing,*. J. Web Sem. 10, p. 1-2.

Heise, A., Naumann, F., 2012. *Integrating open government data with stratosphere for more transparency*, J. Web Sem. 14, p. 45-56.

Hitzler, Pascal, Schärfe, Henrik, 2009, *Conceptual Structures in Practice*, Taylor and Francis Group.

HIS-Building A Roadmap for Health Information Systems Interoperability for Public Health, 2007. Dec.2012 http://www.ihe.net/technical_framework/upload/ihe_phdsc_public_health_white_paper_2007_10_11.pdf.

Hong, Tzung-Pei; Chen, Chun-Hao; Tseng, Vincent S., 2010, *A Summary of Genetic-Fuzzy Data Mining Techniques*, GrC 2010, pp. 28-29.

IHE – Integrating the Healthcare Enterprise, www.ihe.net, Accessed December 2012.

IMS CC – IMS Common Cartridge http://www.imsglobal.org/cc/, Access December 2012.

Jain, Anil K., 2010, *Data clustering: 50 years beyond K-means*. Pattern Recogn. Lett. 31, 8 (June 2010), pp. 651-666. DOI=10.1016/j.patrec.2009.09.011.

Karnstedt, M., Sattler, K-U, Hauswirth M., 2012. *Scalable distributed indexing and query processing over Linked Data*, J. Web Sem. 10, p. 3-32.

Kassahun, A., Beulens, A., Hartog, R., 2006. Providing Author-Defined State Data Storage to Learning Objects, Educational Tech. & Society, 9(2), 19-32.

Kohonen, T., Schroeder, M. R. and Huang, T. S. (Eds.) , 2001. *Self-Organizing Maps* (3rd ed.). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Lassila, O., Swick, R., 1999. *Resource Description Framework (RDF) Model and Syntax Specification*, W3C (World-Wide Web Consortium) http://www.w3.org/TR/1999/REC-rdf-syntax-19990222 , Accessed December 2013.

Le, D. X., Rahayu, J. W., and Pardede, E., 2006. Dynamic approach for integrating web data warehouses. In Proc. of International Conference on Computational Science and Its Applications (ICCSA). 207–216.

LTI-IMS – "Learning Tools Interoperability". *IMS Global Learning Consortium* , Accessed December 2012 http://www.imsglobal.org/toolsinteroperability2.cfm ,

Manolescu, I., Florescu, D., Kossmann, D., 2001. Answering XML Queries on Heterogeneous, VLDB '01 Proceedings of the 27th International Conference on Very Large Data Bases. 241-250.

Mitchell, Tom, 1997. *Machine Learning,* McGraw Hill.

Morris, E., Levine, L., Meyers, C., Place, P., Plakosh, D., 2004. *System of Systems Interoperability (SOSI): Final Report*, Software Engineering Institute, Carnegie Mellon University, Technical Report CMU/SEI-2004-TR-004, http://www.sei.cmu.edu/reports/04tr004.pdf.

Nwana, H.S., 1996. *Software Agents*: An Overview. Knowledge Engineering Review.

OASIS – OASIS ebXML Messaging Services Version 3.0: Part 1, Core Features, 2007. http://docs.oasis-open.org/ebxml-msg/ebms/v3.0/core/ebms_core-3.0-spec.pdf , Accessed December 2012.

Open – The Open Group, Accessed December 2012 http://www.opengroup.org/subjectareas/si.

OpenIoT - Open Internet of Things architecture https://github.com/OpenIotOrg/openiot/wiki/OpenIoT-Architecture , Accessed December 2013.

PESC-CEDS, PESC Common Education Data Standards http://www.pesc.org/interior.php?page_id=208 Accessed December 2012.

Rosetta – Rosettanet Overview: Clusters, Segments, and PIPs (ver 02.13.00), 2011 http://www.rosettanet.org/TheStandards/RosettaNetStandards/PIPOverview/tabid/3482/Default.aspx.

Sattler, K., Geist, I., Schallehn, E., 2005. *Concept-based Querying in Mediator Systems*. VLDB Journal, 14(1), 97–111.

SCORM – Shared Content Object Reference Model http://scorm.com/ , Accesed December 2012.

Shen, W., Norrie, D.H., Barthes, J. A., 2001, *Multi-Agent System for Concurrent Intelligent Design and Manufacturing*, Taylor & Francis.

SIF Specifications, http://www.sifinfo.org/us/sif-specification.asp, Accessed December 2012.

Sutton, R S., Barto, A G, 1998. *Reinforcement Learning*: *An Introduction*, A Bradford Book. The MIT Press, Cambridge, Massachusetts.

Stumme, G., Wille, R., 2000. *Begriffliche Wissensverarbeitung / Conceptual Knowledge Processing*, Springer Verlag.

Şerban, G.**,** Czibula, I. G., Câmpan, A., 2008. *Medical Diagnosis Prediction using Relational Association Rules*, International Conference on Theory and Applications of Mathematics and Informatics (ICTAMI'07), Alba-Iulia, pp. 339-352.

UDEF – Universal Data Element Framework, http://www.opengroup.org/udef/, Acc December 2013.

Vouros, G. A., Papasalouros, A.; Tzonas, K., Valarakos, A G., Kotis, K., Quiané-Ruiz, J-A.; Lamarre, P., Valduriez, P., 2010. *A semantic information system for services and traded resources in Grid e-markets*, Future Generation Comp. Syst. 26(7), p. 916-933.

Walker: E., 2012. *Primer on K-20 Education Interoperability Standards*, Publication of the Software & Information Industry Association.

Witten, I. H., Frank, E., Hall, Mark A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.), Elsevier, ISBN 978-0-12-374856-0.

Weiss, G. E., 1999. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, MIT Press.

Wooldridge, M., Jennings, N. R., Kinny, D., 2000. *The Gaia Methodology for Agent-Oriented Analysis and Design*, Autonomous Agents and Multi-Agent Systems 3, pp. 285-312.

XML standard, http://www.w3.org/TR/xml11/#charsets , accessed November 2012.

Yuan, L., Wilson, S., Cooper, A., Campbell, L. M., 2010. The Future of Interoperability and Standards in Education – System and Process, *CETIS White paper.*

Ziemann, Jörg, 2010. *Architecture of Interoperable Information Systems - An Enterprise Model-Based for Describing and Enacting Collaborative Business Processes*, Logos Verlag, Berlin.

Multimedia Retrieval based on Geometric Distance in Semi-structured Document

Sana Fakhfakh, Mohamed Tmar and Walid Mahdi
Laboratory MIRACL, Institute of Computer Science and Multimedia of Sfax, Sfax University, Sfax, Tunisia
sanafakhfakh@yahoo.fr, {mohamed.tmar, walid.mahdi}@isimsf.rnu.tn

Keywords: Geometric Distance, Image, Multimedia Retrieval, XML Element, Structure.

Abstract: This paper is included on multimedia retrieval in XML document, whose goal is to find relevant multimedia element. In this article, we are particularly interested by studing the impact of various structural factors for image retrieval by our proposed method using a new evidence of source applied to media "image". This method consist to define a geometric distance between XML nodes. Experiments are undertaken into two data sets "INEX 2007" and "ImageCLEF 2010". The obtained results showed effectiveness of our approach.

1 INTRODUCTION

This paper falls under the context of multimedia retrieval in XML documents. The need with this kind of information is justified by quick change of scopes of application which use structural documents (format HTML or XML) what imposes new challenges in the field of search for information. Indeed, nowadays XML document passed a simple tool for exchanging data to a new storage medium. XML document includes textual element and multimedia element such as image, audio and video. These elements are organized according to structure which includes information notably although there is not only one manner to organize contents. However, the choice of structure depends greatly on the context of use of the textual contents.

Mainly in the literature, there are two main classes of approaches in the field of multimedia retrieval: retrieval methods based on multimedia content (MR-content) and multimedia methods to retrieval based on context (MR-Context). The approaches of the multimedia retrieval based on content use specific features of low level according to type of media (Lew, 2006). We can cite for example image retrieval that exploits visual features (the color, texture, forms $\cdots$). These methods have proven effective with media "image" in well defined fields such as medical field this is due to requirement for thorough knowledge of distinctive media. This type of research can be applied to only one type of media in system due to lack of semantic representation in media content.

The approaches of the multimedia retrieval based on context do not depend on type of media in question (Elghazel et al., 2005) (Tjondronegoro et al., 2005). Indeed, these methods rely on information surrounding the multimedia element representing its semantic description. Multimedia retrieval based on textual context is most used, although the structural context remains an obvious source which plays a part paramount in understanding of structured documents.

In this article, we focus on techniques for multimedia retrieval based on textual and structural context in XML documents. This type of document includes textual information and structural constraints. So, XML document cannot be effectively exploited by classical techniques of IR, which regard document as a plane source of information.

The implicit incorporation of multimedia elements in XML documents requires the exploitation of textual context for multimedia retrieval. However, the textual context remains insufficient in most of time. The idea is to calculate the relevancy score of media element based on information from the textual and structural context to answer a specific information needs of user, expressed as query composed of set of keywords.

Let us take for example a image media. If we exploit the image context which is composed by description of its contents such as its title, name, descriptive texts which surround it, title of XML document ... In following figure, we present document extracted from "WIKIPEDIA" encyclopedia describing lion. We notice the existing simultaneous textual and multimedia information. For image retrieval from time after "Pleistocene", we extract information from the tex-

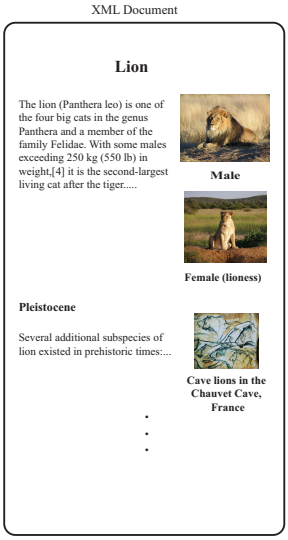


Figure 1: Example of a multimedia object context.

tual description, not from title (figure 1).

In our work, we will be interested by media ”image”. Most existing work in this area uses the information from textual description of image. There are other sources of evidence that were used as visual descriptors, information from link around the image (Hliaoutakis et al., 2006), structure of XML document. To resolve difficulties in mutlimedia retrieval field, you must define adequate source of evidence for representation a multimedia element and defining appropriate indexing model. In this context, we present our structural indexing system combining conceptual information for semi-structured documents dedicated to approximate retrieval data. We begin with an overview of existing work in multimedia retrieval. Then we turn to the presentation of our approach while detailing the preprocessing, extraction of textual and structural and phase calculation relevance of multimedia element in information a better response to needs expressed by user. Finally, we present the results of applying our method on two data sets ”INEX 2007” and ”ImageCLEF 2010”.

2 STATE OF THE ART

The advent of structured documents has caused new problems in information retrieval world, and more specifically in multimedia elements retrieval. These problems are strongly related to nature of these documents that provide the structure as a new source of evidence. Thus, nowadays, XML documents include multimedia elements of different types (audio, video and image)implicitly embedded in the textual elements. These multimedia elements (such as physical objects) do not contain enough information to be able to answer a given query. Therefore, the calculation of relevance score of multimedia element must be linked to textual and structural information provided by other nodes XML (Hliaoutakis et al., 2006).

Several works deal XML document as a flat source of information and ignore the structure of XML documents. In this context, (Schlieder and Holger, 2002) say: ”Ignore the document structure is to ignore its semantics”. Indeed, XML document is used to describe a set of data by a structure that provides a semantic lexicon. Thus, it facilitates the presentation of information in terms of interpretation and exploitation. Replying to this need, new works appear in the field of multimedia retrieval that takes in account the structure as source of relevant information. Existing work in structured retrieval of multimedia elements is decomposed in two classes. The first class includes some works which proceed to adopt some traditional technical of retrieval information as language model. In this context, the team *CWI/UTwente* performs a step of filtering results to keep the fragments containing at least one multimedia element (Tsikrika et al., 2008)(Westerveld et al., 2007). The second class includes the specific work to be structured multimedia retrieval. This class uses the structure as a source of evidence in the process of selection of multimedia elements. As first step, (Kong and Lalmas, 2005) proposed a method which combines structure of XML document (XPath) with the use of links (XLink). This method consist to divide XML document into regions. Each region represent a area of ancestors of the multimedia element. His score is calculated in function of the scores of each region. This method exploits vertical structure only. In a second time, (Torjmen et al., 2010) have used the addition of horizontal structure to the notion of hierarchy. (Torjmen et al., 2010) use a method called ”CBA” (Children, Brothers, Ancestors), which takes into consideration the information carried by the children , brothers and fathers nodes for calculate the relevance of multimedia elements. The authors propose an alternative method ”OntologyLike” which is based on the identification of XML document to ontology. To calculate the similarity between nodes the authors use similarity measures that are mainly based on the number of edges to calculate the distance between nodes. There are other approaches to multimedia retrieval are based on exploitation of links in XML document (Awadi and Torjmen, 2010). This work was improved by proposing a hybrid approach that combines structure with using of links who is consider as semantic links (Aouadi et al., 2012). This method above to divide the docu-

ment into regions according the hierarchical structure and the location of image in document. This factor plays a role in the weighting of links for compute the score of image.

In this paper, we propose a new metric for multimedia retrieval in XML documents which involves the use of geometric distances to calculate the relevance of each node from the multimedia node. This method consists of placing the nodes of XML document in Euclidean space and define each node by a vector of coordinates to calculate then the distance between each pair of nodes. This distance will play a beneficial role in to calculate the score of multimedia element.

3 PROPOSED APPROACH

We focus on techniques for multimedia retrieval based on textual and structural context in XML documents. XML documents cannot be effectively exploited by classical techniques of IR, which regard document as a bog of words. Therefore, the calculation of relevance score of multimedia element must be linked to textual and structural information provided by other nodes XML (Hliaoutakis et al., 2006). Thus, it facilitates the presentation of information in terms of interpretation and exploitation. Replying to this need, we propose a new method in the field of multimedia retrieval that takes into account the structure as a source of evidence and its impact on search performance. We present a new source of evidence dedicated to multimedia retrieval based on the intuition that each textual node contains information that describes semantically a multimedia element. And the participation of each text node in the score of a multimedia element varies with its position in there XML document. To compute the geometric distance, we initially place the nodes of each XML document in an Euclidean space to calculate the coordinates of each node by algorithm 1. Then, we compute the score of a multimedia element depending on the distance between each textual node (Fakhfakh et al., 2013).

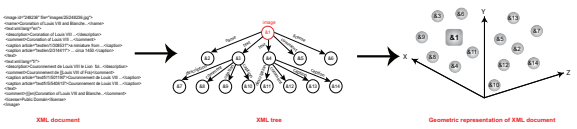


Figure 2: The steps of passing an XML document to geometric representation.

Figure 2 shows the steps of passing an XML document to a geometric representation of the XML elements in a Euclidean space. The first step consist to present a XML document as XML tree to take into account XML document properties.

An XML tree is described by a set of relationships between nodes. Formally an XML tree is a pair $A = (E,R)$ where E is a set of XML elements and $R \subset E^2$, ($(p,q) \in R$ if p is the parent of q) is a set of relations satisfying:

$$\exists! r \in E, \forall q \in E - \{r\}, (r,q) \in R \quad (1)$$

With r is the root of the tree.

$$\forall p \in E - \{r\}, \exists! q \in E, (p,q) \in R \quad (2)$$

Each node has a parent except the root r.

In second step, we will spend to presentation of XML tree in a geometric representation. This step is mainly based on equalities extraction in XML tree according to our proposed hypotheses.

The XML tree representation allowed us to unveil certain relationships of neighboring, brotherhood and offspring. Indeed, the distance d which separate two or more brothers with their common ancestors iteratively is the same. And brothers of the same hierarchical level are equidistant.

These distances are defined according to the relationship of contiguity and semantic similarity between nodes. These distances are not quantized but will be extracted in function of the position of each textual node in XML tree.

All these properties result in: For all $q_i = (x_{i1}, x_{i2} \cdots x_{im})$ and $q_j = (x_{j1}, x_{j2} \cdots x_{jm})$ where Q is a set of vectors in $\mathbb{R}^m$.

- In the same hierarchy, if there are more than two brothers then their adjacent nodes are equidistant:

  property 1

  $$\forall q_i, q_j, q_k \in Q, if\ A_1(q_i) = A_1(q_j) = A_1(q_k)$$
  $$d(q_i,q_j) = d(q_i,q_k)$$

- The distance between any node and its descendants is the same:

  property 2

  $$\forall q_i, q_j, q \in Q, n \in \mathbb{N}, A_n(q_i) = A_n(q_j) = q$$
  $$d(q_i,q) = d(q_j,q)$$

With $\forall n \in \mathbb{N}^*$, we define function A_n by:
$\forall q \in E$,

$$A_n(q) = \begin{cases} \{q\} \text{ if } n = 0 \\ A_{n-1}(p) \text{ if } \exists\ p \in\ E,\ (p,q) \in\ R \text{ and } n > 0 \\ \varnothing \text{ else} \end{cases}$$

From these relationships, we can generate system of equations taking into account for kinship relationships nodes based on hierarchy and adjacency. These relationships are decried by equalities in this order (these equations are only examples):

$$\begin{aligned} d(n_1,n_2) &= d(n_1,n_3) \\ d(n_1,n_2) &= d(n_1,n_4) \\ d(n_1,n_7) &= d(n_1,n_8) \\ d(n_1,n_7) &= d(n_1,n_9) \end{aligned}$$

These distances are defined according to the relationship of contiguity and semantic similarity between nodes. They are not quantized but will be extracted in function of the position of each textual node in the XML tree. The resulting system is nonlinear, its resolution requires the use of an approximate resolution iteratively method where we used iterative solution method (see Algorithm 1).

Algorithm 1: Resolution algorithm approximate nonlinear system of equations.

Require: $(Q=(q_1,q_2...q_{|Q|}),R)$:an XML tree as q_i=(q_{i1},q_{i2}...q_{im}) $\forall i \in [1,|Q|]$
  m:dimension
  for $(i,j) \in [1,|Q|]^2$ **do**
    $q_{ij} \leftarrow$ random value
  end for
  $Q_1 \leftarrow (q_1,q_2...q_{|Q|})$
  repeat
    $P \leftarrow Q_1$
    for $(i,j) \in [1,|Q|]^2$ **do**
      $Q_2 \leftarrow (q_1,q_2...q_{i-1},q_i+d_j(1),q_{i+1}\cdots q_{|Q|})$
      $Q_3 \leftarrow (q_1,q_2...q_{i-1},q_i+d_j(\varepsilon),q_{i+1}\cdots q_{|Q|})$
      $Q_4 \leftarrow (q_1,q_2...q_{i-1},q_i+d_j(1-\varepsilon),q_{i+1}\cdots q_{|Q|})$
      $t \leftarrow 0$
      while $error(Q_1) > error(Q_2) > error(Q_3) > error(Q_4)$ **do**
        $Q_4 = (q_1,q_2...q_{i-1},q_i+2^t d_j(1),q_{i+1}\cdots q_{|Q|})$
        t=t+1
      end while
      $t \leftarrow 0$
      while $error(Q_1) < error(Q_2) < error(Q_3) < error(Q_4)$ **do**
        $Q_1 = (q_1,q_2...q_{i-1},q_i-2^t d_j(1),q_{i+1}\cdots q_{|Q|})$
        t=t+1
      end while
      while $|error(Q_1)-error(Q_2)| > \varepsilon$ **do**
        $Q_5 \leftarrow \frac{Q_1+Q_2}{2}$
        let $Q_5=(p_1,p_2...p_{|Q|})$
        if $error(p_1,p_2...p_{i-1},p_i - d_j(\varepsilon),p_{i+1}\cdots p_{|Q|}) > error(p_1,p_2...p_{i-1},p_i+d_j(\varepsilon),p_{i+1}\cdots p_{|Q|})$ **then**
          $Q_1 \leftarrow Q_5$
        else
          $Q_2 \leftarrow Q_5$
        end if
      end while
    end for
  until $P = Q_1$

The process begins by assigning to each XML node a random vector it. Tries to improve the coordinate values of each node according to an error value (the sum of the squared deviations). At each iteration, the coordinates are improved together with the minimization of this error. The algorithm stops when the error reaches its minimum value (no improvement is possible). Let Q the set of vectors obtained at a given iteration during the running of the algorithm, the error is defined by:

$$\begin{aligned} error(Q) = & \sum_{\substack{q_i,q_j,q_k \in Q, \\ A_1(q_i)=A_1(q_j)=A_1(q_k)}} (d(q_i,q_j)-d(q_i,q_k))^2 \\ & + \sum_{\substack{q_i,q_j,q \in Q, n \in \mathbb{N}, \\ A_n(q_i)=A_n(q_j)=q}} (d(q_i,q)-d(q_j,q))^2 \end{aligned}$$

Where m is the dimension of the Euclidean space and $\forall v \in \mathbb{R}$, $D_j(v)=(d_1,d_2\cdots d_m)$ is such as:

$$d_k = \begin{cases} 0 \; if \; k \neq j \\ v \; otherwise \end{cases}$$

3.1 Indexing System

We propose a indexing system $MXS-index$ composed by two party: party of textual indexing and party of structural indexing. In first party, our approach uses NLP (Natural Language Processing) techniques to extract the candidate XML nodes of the resulting indexing. The weight of these nodes is depending on the frequency of each of these terms and the number of elements in the corpus according to the number of elements containing the term. In Second party, we built structural index using information extract from XML tree and geometric metric. Each XML node will presented by characteristic vector. We start by extract geometric proprieties. And we compute coordinates of each XML nodes. This party is accompanied by generating XML data model which processes ancestor, descendant and proximity relationships (figure 3).

Figure 4 schematize the process of textual and structural indexing XML documents with our indexing system. Well as the transition of XML document as a tree presentation to geometric presentation in Euclidean space.

3.2 Multimedia Element

A multimedia element (eg *image*) does not contain textual content. Its score is based on textual nodes in its neighborhood. The transition from the XML tree structure representation of elements in an Euclidean space, where we exploit the dissimilarity distances separating a multimedia node and other textual nodes, is performed by extracting the equations satisfying the properties defined earlier and the application of algorithm 1. To calculate the distance between a node n and multimedia element H, we calculate the Euclidean distance between their respective feature vec-

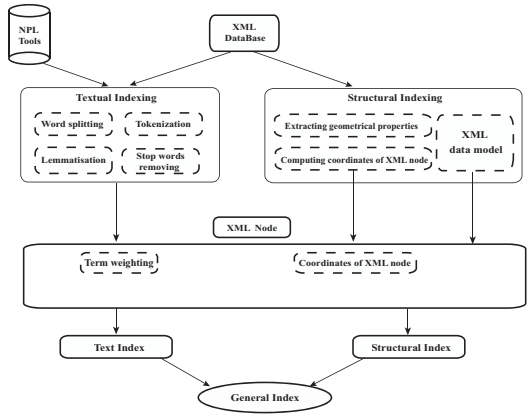


Figure 3: Architecture of our indexing model $MXS-index$.

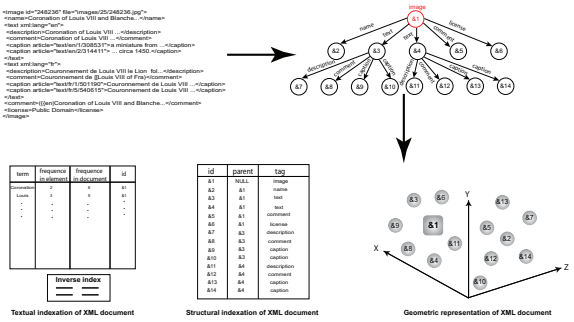


Figure 4: Treatment process of XML document.

tors q_n and q_H:

$$dist(n,H) = \sqrt{\sum_{i=1}^{m}(q_n - q_H)^2} \quad (3)$$

With m is the dimension of the Euclidean space. q_n is defined by: q_n=(xn_{i1}, xn_{i2} ... xn_{im}) with xn are the vector characteristics of node n. And q_H is defined by: q_H =(xH_{i1}, xH_{i2} ... xH_{im} with xH represent the coordinates compose the vector characteristics of a node H. We calculate the score for each textual node depending on the frequency of each term (tf) and the number of elements in the corpus according to the number of elements containing the term (idf). A textual node is presented by: $n = (n_1, n_2 \cdots n_{|v|})$ where n_i is the weight of the term t_i, v is the set of indexing terms:

$$n_i = tf(t_i,n) \times idf(t_i) \quad (4)$$

With

$$idf(t_i) = log(\frac{N}{N_i}) \quad (5)$$

Where N is the total number of XML elements in the corpus, N_i is the number of elements that contain the term t_i and $tf(t_i,n)$ is the frequency of the term t_i in node n. The score of textual node depends on the weight of each indexing term. A query is made by the list $v = (v_1, v_2 \cdots v_{|v|})$ where $v_i \in \{0,1\}$ (0:not exist, 1:exist) according membership t_i at the query. The score of textual node n for the query q is defined by:

$$rsv(q,n) = q \times n^T = \sum_{i=1}^{|V|} q_i \times n_i \quad (6)$$

Where μ is the set of textual elements. The score of multimedia node H is defined by:

$$rsv(q,H) = \sum_{n \in \mu} \frac{rsv(q,n)}{dist(n,H)} \quad (7)$$

With $dist(n,H)$ is the distance between the feature vectors corresponding to the nodes n and H. This equation leads to assign the importance of contribution of all nodes in computing the score of multimedia element that shows its beneficial impact in multimedia retrieval.

4 EVALUATION AND RESULTS

We evaluate our system into two databases extracted from two collections : INEX 2007 (Initiative for the Evaluation of XML Retrieval) Ad Hoc task (Fuhr et al., 2007) and ImageCLEF 2010 Wikipedia image retrieval task (Popescu et al., 2010). These databases are composed by XML documents extracted from Wikipedia. The evaluation results show that this method provides a MAP which is equal to 0.2102 as MAP with using "ImageCLEF 2010" collection. The result has been improved significantly with the "INEX 2007" collection to 0.3102 as MAP. This increase is due to nature of "INEX 2007" collection who includes XML documents with heterogeneous structure. So in "INEX 2007" collection we find documents with high depth. This factor highlights structural information and amplifies effect textual information based on computed distances . For against, our system is more stable with "ImageCLEF 2010" collection, this is due to rapid convergence of results (algorithm 1). With our measure, we have shown that combined use of textual and structural context can properly determine the relevance of multimedia element, and the structure plays a primordial role in multimedia retrieval.

Table 1: Results of the impact our approach on INEX 2007 and ImageCLEF 2010 based in *MAP*(Mean Average Precision).

Company	**INEX 2007**	**ImageCLEF 2010**
Task	Collection XML Ad Hoc	Wikipedia Retrieval
Number of XML document	659,388	237,434
Number of image	246,730	237,434
Topics	19	70
MAP	0.3102	0.2572

5 CONCLUSIONS

In this paper, we propose a novel approach for multimedia retrieval in XML documents. This method consist to calculate the score of element mulltimedia according the textual context provided by nodes in proximity and structural context from distance between nodes and multimedia element. Experiments shows the interest of our method on INEX 2007 and ImageCLEF 2010 collections. In the future, we want to exploit another factor to calculate the relevance of multimedia element such as the title of image, the weighting of the links in XML document ... As well as another source of evidence as visual descriptors and the study parameters combination of using of structural, textual and visual context.

REFERENCES

Aouadi, H., Torjmen-Khemakhem, M., and Jemaa, M. B. (2012). Combination of document structure and links for multimedia object retrieval. *Journal of Information Science*, 38(5):442–458.

Awadi, H. and Torjmen, M. (2010). Exploitation des liens pour la recherche d'images dans des documents xml.

Elghazel, H., Idrissi, K., Baskurt, A., and Ben Amar, C. (2005). Approche textuelle pour la recherche d'image. In *3rd International Conference on Sciences of Electronic, Technologies of Information and Telecommunications SETIT 2005*.

Fakhfakh, S., Tmar, M., and Mahdi, W. (2013). A new metric for multimedia retrieval in structured documents. In *ICEIS (2)*, pages 240–247.

Fuhr, N., Kamps, J., Lalmas, M., Malik, S., and Trotman, A. (2007). Overview of the inex 2007 ad hoc track. In *INEX*, pages 1–23.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., and Milios, E. (2006). Information retrieval by semantic similarity. In *Intern. Journal on Semantic Web and Information Systems (IJSWIS).Special Issue of Multimedia Semantics*, pages 55–73.

Kong, Z. and Lalmas, M. (2005). Xml multimedia retrieval. In *SPIRE*, pages 218–223.

Lew, M. S. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl*, 2:1–19.

Popescu, A., Tsikrika, T., and Kludas, J. (2010). Overview of the wikipedia retrieval task at imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

Schlieder, T. and Holger, M. (2002). Querying and ranking xml documents. *Journal of the American Society for Information Science and Technology*, 53:489–503.

Tjondronegoro, D., Zhang, J., Gu, J., Nguyen, A., and Geva, S. (2005). Integrating text retrieval and image retrieval in xml document searching. In *INEX*, pages 511–524.

Torjmen, M., Pinel-Sauvagnat, K., and Boughanem, M. (2010). Using textual and structural context for searching multimedia elements. *IJBIDM*, 5(4):323–352.

Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D., and de, A. P. V. (2008). Structured document retrieval, multimedia retrieval, and entity ranking using pf/tijah. In *6th Initiative on the Evaluation of XML Retrieval, INEX 2007*, volume 4862 of *Lecture Notes in Computer Science*, pages 306–320, London. Springer Verlag.

Westerveld, T., Rode, H., van, R. O., Hiemstra, D., Ramirez, G., Mihajlovic, V., and de, A. V. (2007). Evaluating structured information retrieval and multimedia retrieval using pf/tijah. In Fuhr, N., Lalmas, M., and Trotman, A., editors, *Comparative Evaluation of XML Information Retrieval Systems*, volume 4518 of *Lecture Notes in Computer Science*, pages 104–114, Berlin, Germany. Springer Verlag.

Extensible Data Management Architecture for Smart Campus Applications
A Crowdsourcing based Solution

Attila Adamkó and Lajos Kollár
Department of Information Technology, Faculty of Informatics, University of Debrecen, Debrecen, Hungary
{*adamkoa, kollarl*}*@inf.unideb.hu*

Keywords: Smart Campus, Crowdsourcing, Participatory Sensing, Data Integration.

Abstract: The technological advancements that have occurred during the past decade in various domains, including sensors, wireless communications, location positioning technologies and the web, allow the collection of a wide range of data. Possible sources of that data include intelligent devices (smartphones, tablets, etc.) containing various sensors, Web pages, and social networking sites. Collected data are subject to analysis (using data mining or pattern recognition approaches, for instance) and after processing new content might be inferred. This is a value-added service that can itself be used as a data source. In this paper, we use our University Campus as an example for establishing a data management architecture that integrates into a more general, extensible publish/subscribe based model of crowdsourced applications.

1 INTRODUCTION

The recent rise in ubiquitous computing (ubicomp) and supporting devices resulted in lots of new applications that are able to exploit the advantages of this approach. According to (Luo, 2012), "ubicomp is a post-desktop model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities". The collected data from the built-in sensors of smartphones, tablets, phablets, and embedded devices let the applications based on participatory sensing become more and more popular and useful. Moreover, in a real-life environment we need additional (user- or application-generated) data sources in order to provide valuable services.

A Smart Campus is a good prospect of applying participatory sensing combined with other data and event sources and due to the number of people studying, working or even living at a Campus and considering their commitment of using novel technologies and applications, it is an ideal choice for developing value-added services based on crowdsensing.

This paper is organized as follows. Section 2 describes the context of our research and explains the reasons why we selected Smart Campus for our focus. Key challenges and our research goals are also defined here. Section 3 briefly describes the related work. The main contribution of the paper, namely, the supporting architecture is discussed in Section 4 while Section 5 contains some concluding remarks.

2 SMART CAMPUS AS A PART OF A SMART CITY

The context of this research is a project entitled Future Internet Research, Services and Technology (FIRST[1]) has been started in 2012 by University of Debrecen and its consortial partners. The research objectives of the project include to face the challenges posed by the Internet, to carry on investigations covering basic theoretical questions, network modelling examinations, the reassessing of network architecture, content management, and the creation of an intelligent application platform.

The aims of one of its subprojects, Future Internet for Smart City Applications, are threefold:

1. to establish an open, mobile crowdsensing application platform for smart city applications,
2. to provide data management and knowledge discovery solutions for them, and
3. to develop prototype applications based on the

[1] http://first.tamop422.unideb.hu/About-the-project.html?language=en

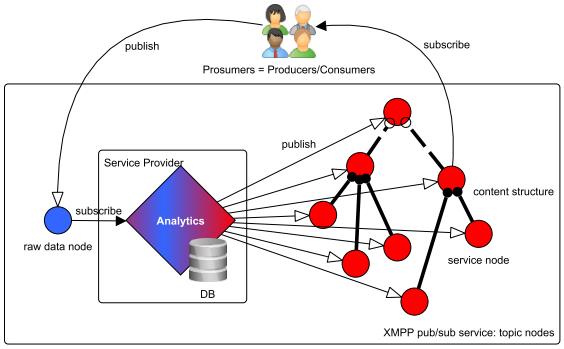


Figure 1: XMPP-based publish/subscribe architecture (Szabó and Farkas, 2013a).

common platform combining the data management solutions with real-time analytics.

The open platform has already been established (Szabó and Farkas, 2013a; Szabó and Farkas, 2013b; Szabó et al., 2013). Figure 1 shows how the model can directly be mapped onto the Extensible Messaging and Presence Protocol's (XMPP) publish/subscribe (pubsub) service (Saint-Andre, 2011).

During our work to design data management infrastructure, we started to investigate the field of Smart Campuses instead of Smart Cities (fortunately, this also fitted into the overall project). The reason for it was that there are lots of similarities between the challenges of developing Smart Campus and Smart City applications (the corresponding terms used by Figure 2 and (Szabó et al., 2013) are given inside parentheses):

Lots of Potential Users (Consumers)
University of Debrecen has as much students and employees as the number of inhabitants of a medium-sized Hungarian city. These users are consuming the provided services.

High Variety of Data Sources (Producers)
Information in various formats coming from multiple sources should be integrated in both cases (e.g., timetable, academic calendar, information on consultation dates and times in case of Smart Campus, or energy consumption, traffic info, etc. for Smart Cities). Social media sites and geolocation sensors are examples of sources that are common in both domains.

Need for Value-added Services (Service Providers)
Service Providers give added value to the crowd-collected raw data. Basically we can say that this is why someone would like to use the applications (e.g., for being notified about a room change or finding a jam-free route between two points).

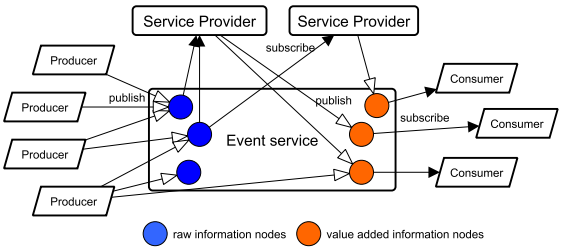


Figure 2: Crowdsourcing model based on a pubsub scheme (Szabó et al., 2013).

Examples of such value-added services include but not limited to:

- get notified when one of my favourite bands (based on my listening history from last.fm or Google Play Music) will play in a concert venue of the Campus,
- get notified when an event's (class, concert, etc.) date and time is modified,
- based on the location data provided by some of my friends I can realize that they are going to have a lunch so if I am in a hurry I can join them,
- when preparing for a consultation with an instructor, get notified if too many students plan to visit the instructor's office hour at the same time, and assist of reorganizing my schedule based on my task list.

We have a much longer list describing various use cases for Smart Campus applications but since the focus of this paper is not on the applications but the underlying data management architecture, we omit the details of use cases.

An advantage of building a prototype application on the Smart Campus field over a Smart City is that young people's (especially students') willingness to use the applications and provide valuable feedback on them is much bigger than elders' therefore it seemed to be a good idea to test and evaluate the underlying data management architecture. On the other hand, participatory sensing can successfully be applied as students actively use handheld devices and social media (presumably they are much more active than city inhabitants are). Our aim is to build services that can be used to infer some patterns regarding the operation of the community, and the derived information can be utilized as a part of a new service offered for the community after applying appropriate analytics.

Smart Campus applications should fit into this framework so they need to follow an XMPP-based pubsub approach with crowdsourcing. However, in order to achieve platform independence, we decided to encapsulate services into Web services which allow

to develop consumer apps on any of the most popular mobile platforms (Android, iOS, Windows Phone) along with traditional Web applications.

2.1 Challenges

A Smart Campus environment has lots of various (and most importantly, heterogeneous) data sources including the following:

- an Education Administration System called Neptun that contains information on course enrollments, timetable information of courses, exam dates and times, etc.,
- faculty members offering office hours, consultations, etc.,
- Education Offices of the various faculties offering office hours,
- Student Governments organizing events for students,
- the menus of the canteens located at the Campus,
- geolocation (e.g., GPS), WiFi or some other sensor data collected by smartphones or similar devices,
- data gathered by environmental and building sensors (temperature, humidity, air pressure, air pollution, etc.),
- a Library Information System that is able to tell whether a given book is available or not,
- social media sites (like Facebook or Google+) containing information on friends and ranges of interests of a person,
- professional sites (like LinkedIn) holding data on work experience and professional achievements (however, this is not necessarily the most important data source from Smart Campus perspective),
- bibliographic databases (like Google Scholar, DBLP or Scopus) that provide information of published journal articles or conference papers of researchers,
- event hosts of actually any events (like public lectures, concerts, exhibitions or whatever users might be interested in), and, which is essential,
- the crowd itself with the added value of the capability of generating content that is interesting for a set of people (or, to be more precise, consumers).

Of course, this is not an exhaustive list, these have only been listed in order to demonstrate how diversified sources of data can be used in a complex application. Valuable applications typically require integration of some data originating from several of these sources. Therefore, our main challenge is to create an architecture which allows the access of information from existing sources while also easing the addition of new sources in a seamless manner.

Some of those data can be gathered in an automated way (like sensor data), some of them may require manual interaction (e.g., canteens' menus or instructors' office hours); some of those data sources offer Application Programming Interfaces (APIs) to provide access to data (social media sites, for instance) while others do not have APIs so some web spiders are required to parse the data; some of the data sources provide built-in notification mechanisms (e.g., an event feed of a social network site) while others do not (for example, adding new office hours or changing the daily menu).

Considering the amount of data originating from those sources, we can state that a big data management solution is required. The frequency of arriving data elements can significantly differ, some of them (e.g., inserting a new article into a bibliographic database or offer office hours for the forthcoming semester) might be quite rare while others (especially when collecting location data) can be very frequent.

This will be further discussed in Section 4.

2.2 Our Goal

An architecture for data management and knowledge discovery is needed that fits into the framework shown in Figure 1. This architecture should address the challenges described in Section 2.1 and be as generic as it can be in order to integrate a high variety of data sources regardless of how the data is processed. Extensibility is a key concern: the design should allow to add new sources and define their processing in an easy way.

Besides setting our goals, it might be useful to discuss what are the non-goals. On one hand, discussions on the use of XMPP will not be provided as the overall framework was given. On the other hand, the development of value-added services may (and will) require data mining activities (including understanding the semantics of data in order to eliminate duplications), as well, however, they are completely out of scope of this paper as we here focus only on the data management platform that can (and should) later be complemented with on-line data analytics solutions which are currently under development (by some of our colleagues). Data security is an important issue that needs to be addressed in the technology stack, however, our opinion is that it must be introduced at a higher level than the data management level, therefore we omit the discussions regarding that field.

3 RELATED WORK

There have been innovative steps in Smart City development over the past few years. It became essential to develop a platform that will aggregate all the available (related) information and will orchestrate it in the best way as possible, towards meeting the defined goals.

3.1 Smart Campus

There is an even growing literature for Smart Campuses which are related to data integration and real-time processing of massive heterogeneous data sources. An application prototype using semantic technologies is discussed in (Boran et al., 2011). The idea behind this approach is to include semantic information using ontologies with OWL and SPARQL. Another approach (Valkanas and Gunopulos, 2013) deals with event detection from real-time data using heterogeneous sources.That experiment shows that information extracted form social networks and sensor data can be used for identifying events and they show how affected users (that are physically close to the event source) will be notified based on their actual geolocation.

The high majority of related literature approaches Smart Campuses from an Internet of Things (IoT) point of view, having a focus on the physical campus infrastructure including sensor networks built into buildings, physical entities and places (e.g., parking). This is promising from the energy efficiency point of view but basically provides the same challenges as Smart City projects do. As we were looking for a bit different kind of challenges, we decided to turn to the social factors more than infrastructural. Besides finding the need for integration of data coming from sensor networks very important we think that non-infrastructural aspects should play an equally important role in a life of a Campus. Therefore, people living, studying and working on the Campus should heavily be involved so we feel that exploitation of the crowd's power has a crucial effect on the success of any Smart Campus-related projects.

3.2 Crowdsourcing

The term crowdsourcing is not easy to define. It is a hybrid of "crowd" and "outsourcing", meaning that a given task is assigned to a number of people working on it explicitly. This approach is too restricting, as it is coined in (Doan et al., 2011), therefore they give the following definition: "a system is a crowdsourcing (CS) system if it enlists a crowd of humans to help solve a problem defined by the system owners, and if in doing so, it addresses the following four fundamental challenges: How to recruit and retain users? What contributions can users make? How to combine user contributions to solve the target problem? How to evaluate users and their contributions?"

Starting from these challenges, (Doan et al., 2011) also gives a classification of CS systems. Nine dimensions are identified but here we only concentrate on those that are relevant for us. Based on the nature of collaboration, CS systems can be either explicit or implicit. Both are important for us as explicit systems include evaluating (rate courses, meals, etc.), sharing (location info, files, etc.) and networking (adding new friends or classmates) while implicit systems might be either standalone (when event prediction is based on the history of users' activities) or piggyback (for example, based on the trajectories of a user's movement, predictions can be made). In the context of Smart Campus applications, for the first fundamental challenge (recruiting users) only voluntary participation is possible. For more classes of CS systems with detailed explanations, see (Doan et al., 2011).

3.3 Graph Databases

Graph databases are gaining popularity in the past few years as they provide adjacency structures for data elements without having any indices (i.e., data elements contain direct pointers showing their adjacent elements). The three basic building blocks of graph databases are *nodes* (representing entities), *properties* (pertinent information related to nodes) and *edges* (connecting nodes to nodes using directed arcs). According to the Property Graph Model (Robinson et al., 2013, p. 4), nodes contain properties that are key—value pairs, and edges (a.k.a. relationships) can also have properties. This organization of data allows them to be processed using the well-known graph algorithms.

3.3.1 Advantages of Graph Databases

According to (Robinson et al., 2013), graph databases offer three key advantages:

Performance. Graph databases are easy to scale since queries are localized to a portion of the graph only therefore query execution time is proportional only to the size of the affected sub-graph rather than the size of the overall graph.

Flexibility. Graphs are "naturally additive, meaning we can add new kinds of relationships, new nodes, and new subgraphs to an existing structure without disturbing existing queries and application

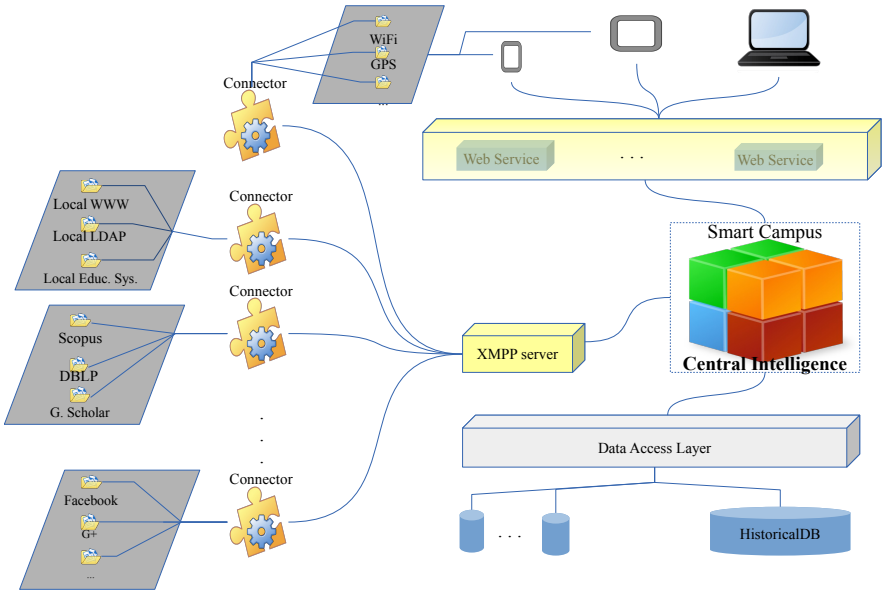


Figure 3: Our extensible architecture.

functionality. These things have generally positive implications for developer productivity and project risk. Because of the graph models flexibility, we do not have to model our domain in exhaustive detail ahead of timea practice that is all but foolhardy in the face of changing business requirements."

Agility. Since the graph data model has a schema-free nature, graph databases suit well for iterative incremental style software development practices. Even the evolution of the data model can be performed in small, agile steps which is a huge advantage when talking about data integration. One can easily add new properties caused by emerging data sources to the existing data model without the need of performing complex schema evolution steps (which are, in fact, probably the biggest challenge when using relational databases).

The graph data model fits well to the data structure of many of today's popular social networking sites inluding Facebook, Twitter, or LinkedIn. It is a definitive advantage over relational databases, key–value stores, and XML databases when our goal is to integrate data from such sources.

4 EXTENSIBLE DATA MANAGEMENT ARCHITECTURE

One of the primary objectives of our system is to capture continuous data streams arriving from different sources. As we discussed in Section 1, Extensible Messaging and Presence Protocol (XMPP) protocol, defining a standardized way of event-based messaging, has been selected as the underlying communication protocol (Szabó and Farkas, 2013b), due to its extensibility and publish/subscribe model. Based on the heterogeneity of the data we should prepare proper input channels for them.

Our proposed architecture (see Figure 3) introduces the term *Connector* which is a node responsible for collecting data from given sources and send updates to the Smart Campus. The notification—and data transfer—is achieved by sending a special XMPP message to the Smart Campus Collector Interface (as shown in Figure 4). Using this information it can be decided whether to approve or decline any updates based on the gathered information. The repository update is declined when a Connector sends a redundant piece of information, for instance, when the Research Connector integrating several sources like DBLP, Scopus, Google Scholar, etc. realizes that it collects the same piece of data that has already been found earlier.

The Connectors in our proposal are crowdsourced compound entities without any deep business logic. They containing the necessary parser or API implementation for the proper sources for a given field—like a parser for Google Scholar and an API for Scopus. It can be imagined as an open box which can be extended (for existing domains) or created and attached to the system as a brand new one—as a new and previously not existed field. This is the power of

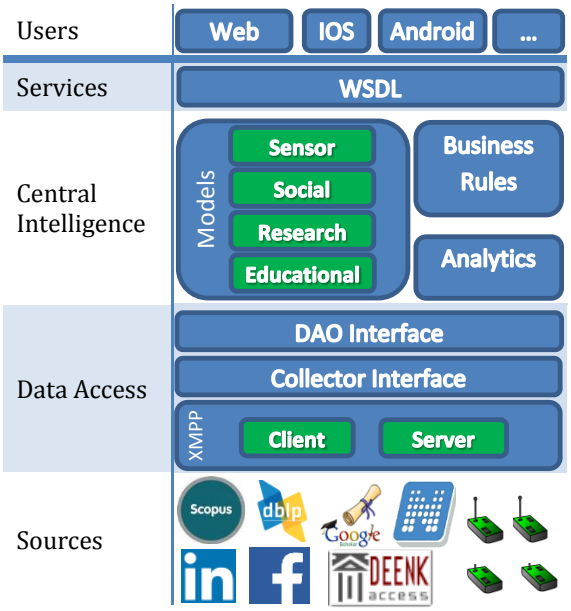


Figure 4: Smart Campus layers.

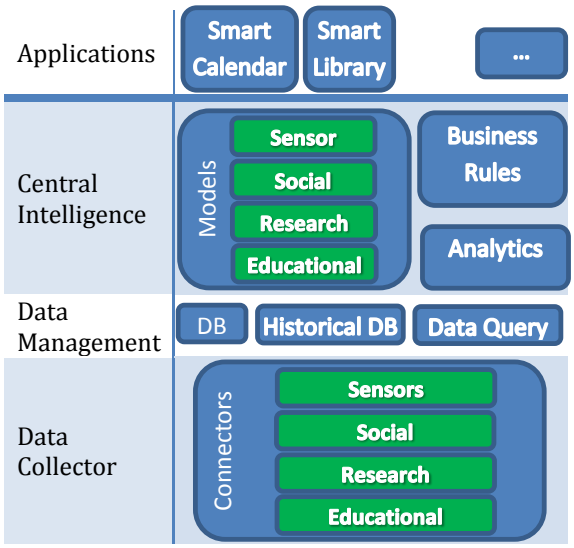


Figure 5: Smart Campus services.

the architecture because it can be done by the crowd. The only requirement is to prepare the Connector for the proper XMPP message structure. If the XMPP server has not got the required message type the extension mechanism of the XMPP comes into the picture. It can be created, inserted and published to the server.

Figure 4 shows a layered view of our Smart Campus architecture. The bottom layer shows the various (and incomplete) sources that are used for gathering data.

The Data Access Layer also plays an important role in our architecture. It provides quick access to the more-or-less statical personal data and to the "fresh" real-time data. It means, the frequently changing data (like sensor data, news feed) can be easily accessed only for a limited time, e.g., three or six months and after all it goes to the HistoricalDB part. This separation can speed up the queries targeted for "fresh" data because the data store cannot grow to a huge one slowing down all the searches. However, the other part, the historical one introduces new challenges as well. At one hand, efficient handling of big data. On the other hand, proper support for the Analytics module to find new and relevant information from that huge amount of data. Since our original goal was to design a data management platform, development of the Analytics module has been postponed until a subsequent iteration.

The Smart Campus Central Intelligence (SCCI) component in our architecture provides an interface between the information sources including both the incoming events (XMPP server) and the information stored in the database and the Web services layer (Figure 4). A service-oriented view of our architecture is shown in Figure 5. The Data Collector layer uses Connectors to connect sources for receiving data. The collected data is then used to populate the databases of the Data Management layer. Applications that are created based on Web services are at the top of the view.

The underlying data model for Smart Campus also highlights challenges. We need to face structured and unstructured data as well. Moreover, it needs to be an extensible model which leads us to a semi-structured model where we know it does not fit into a uniform, one-size-fits-all, rigid relational schema. Faced with the need to generate ever-greater insight and value-added services we turned to graph technologies to tackle the complexity. Besides increasing volume of data, another force which has to be contended with is that over time, data will become more and more unstructured. As data volumes grow, we trade insight for uniformity; the more data we gather about a group of entities, the more that data is likely to be semi-structured but this the way what we originally imagine, an extensible system. But insight and end-user value do not simply result from collecting large volume and variation from data. Insight depends on understanding the relationships between entities, so we need to map how the entities in our domain are connected. The key thing about such a model is that it makes relations first-class citizens of the data, rather than treating them as only metadata. With this approach we established a data store which can be extended in an easy way. When a new—and previously

not existed—source is attached to the system we only need to add the new nodes and proper edges to the database.

5 CONCLUSIONS

In this paper we briefly described a data management architecture for Smart Campus applications. This architecture fits well into the more general publish/subscribe based architecture of Smart City and Smart Campus applications. The developed architecture is extensible with new data sources (appropriate Connectors need to be developed when adding a new source) providing the capability of integration of heterogeneous data. Our future plans include to develop more applications based on that architecture along with the addition of the Analytics module.

ACKNOWLEDGEMENTS

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

REFERENCES

Boran, A., Bedini, I., Matheus, C. J., Patel-Schneider, P. F., and Keeney, J. (2011). A smart campus prototype for demonstrating the semantic integration of heterogeneous data. In Rudolph, S. and Gutierrez, C., editors, *RR*, volume 6902 of *Lecture Notes in Computer Science*, pages 238–243. Springer.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96.

Luo, X. (2012). Social network-based media sharing in the ubiquitous environment: Technologies and applications. In Abraham, A. and Hassanien, A.-E., editors, *Computational Social Networks*, pages 349–366. Springer London.

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O'Reilly Media.

Saint-Andre, P. (2011). RFC 6120: Extensible Messaging and Presence Protocol (XMPP): Core.

Szabó, R., Farkas, K., Ispány, M., Benczúr, A. A., Bátfai, N., Jeszenszky, P., Laki, S., Vágner, A., Kollár, L., Sidló, C., Besenczi, R., Smajda, M., Kövér, G., Szincsák, T., Kádek, T., Kósa, M., Adamkó, A., Lendák, I., Wiandt, B., Tomás, T., Nagy, Á., and Fehér, G. (2013). Framework for smart city applications based on participatory sensing. In *Proceedings of the* 4th *IEEE International Conference on Cognitive Infocommunications*. to appear.

Szabó, R. L. and Farkas, K. (2013a). A publish-subscribe scheme based open architecture for crowd-sourcing. In Bauschert, T., editor, *EUNICE*, volume 8115 of *Lecture Notes in Computer Science*, pages 287–291. Springer.

Szabó, R. L. and Farkas, K. (2013b). Publish/subscribe communication for crowd-sourcing based smart city applications. In *Proceedings in Conference of Informatics and Management Sciences*, pages 314–318.

Valkanas, G. and Gunopulos, D. (2013). Event detection from social media data. *IEEE Data Eng. Bull.*, 36(3):51–58.

Mobile Multi-modal Learning Facilitated through Twitter
A Case Study of Twitter Phenomenology with Graduate Level Nurses

Elizabeth M. La Rue[1], Lauren Terhorst[2] and Kathleen G. McCafferty[3]
[1]*School of Nursing, University of Pittsburgh, Victoria St. Pittsburgh, PA, U.S.A.*
[2]*School of Health and Rehabilitation Services, University of Pittsburgh, Pittsburgh, PA, U.S.A.*
[3]*Applied Technology & Process Consultant, Pittsburgh, PA, U.S.A.*
{eml17, lat15}@pitt.edu, kgmccafferty@gmail.com

Keywords: Mobile Learning, Educational Technology, Teaching Methods, Computer Uses in Education, Higher Education, Nursing Informatics.

Abstract: Adapting academic course content to a mobile world continues to evolve as technology changes. A conceptual deployment of Twitter in a graduate level nursing informatics course is discussed along with how tweeting may fit learning styles, human working memory capacity and reduce the reliance on content management systems to facilitate an online course. Mechanisms for using Twitter to deliver course content and pollinate student interaction while maintaining private individual Twitter accounts for the faculty member and students are discussed.

1 INTRODUCTION

Few would argue that innovative technology and the pervasiveness of information tools into everyday life have changed the speed and flow of communication. Any broadcast message on a social media mechanism will likely receive some form of human feedback within seconds – from anywhere in the world. This availability of 'free' global information has changed our behaviors in accessing information and is influencing a change in the ways in which we learn. We still use our brain, eyes, ears, and written word but our attention span, vocabulary, and tolerance for the mundane hasdecreased in every aspect.

Over the last few yearsin my role as an assistant professor, I have been trying to be an innovator, however, I have come to realize that nothing is ever 'new' and that I will never be first with the idea. My strengths lie in: a) being aware, testing and trying the new tools; b) application of ideas in an educational context; and c)sharing my innovative concepts of repurposing the existing tools through written word.

2 BACKGROUND

Classroom content management systems (CMSs) have ruled the domain of online learning over the past few years as the popular tool to use when teaching online. But the reality is that the CMSs do not teach, they simply assist users in organizing content while requiring significant networkbandwidth and multiple navigationclicks for the users, i.e. extensive user time. They hold information placed in the provided structured network interface. The teaching aspect of online learning is mostly self-led while instructors moderate student discussion boards and grade assignments. After moderating multiple sessions of an online course, this instructor has encountered boredom with the CMS tool and the students have expressed disdain for the online discussion. To recapture interaction with students and interest in the course, change in the facilitation of the course must occur.

Social media tools such as Facebook, Twitter, LinkedIn and communication technologies such as short message service (SMS), Instagram, and Flipboard have revolutionized the way the world interacts. Employing these technologies into an online class may have the ability to unite the students into an online learning network resource group and make their course work and learning a

daily dynamic in their life routine.Not only would they have a 'cool' way to learn and communicate, they could become first adopters of the technology in their respective peer groups. However, using all of these tools in one course would overwhelm the student as well as the instructor. Moderation is required. After thoughtful deliberation and experimentation, Twitter was selected as the most appropriate learning tool for the current proposed intervention.

Twitter is a tool that limits communication to 140 characters. One could say, 'chunk of information' if making a comparison to Miller's 1956 Chunking Theory (Miller, 1956). This chunk of information would be considered a tweet from theTwitter software. Miller proposed that individualshave the ability to retain 7 numbers, plus or minus 2, in their working memory. While comparing 7 digits with or without ±2 digits to 140 characters is quite different, there are knownmemory advantages for the characters. The 140 characters create words that hold the potential to tell stories. Storytelling is a praised teaching methodology known to stimulate student interest and increase information recall (McKee and Fryer, 2003, Dewey, 1933).

Within the story told by a tweet the use of the hashtag incorporates a symbol. Saloman found that symbols have the ability to reduce the mental load and processing power, increase the quantity of the message being delivered and ease recall in individuals (Saloman, 1979).

Symbols can also be referred to as pictures, graphics, and include color to depict a message in Twitter to build a tweet. With the inclusion of pictures and symbols in a tweet, Twitter software actually increases the quantity of information transmitted in only a 140 characters. The adage of, 'A picture is worth a thousand words' is further strengthened with research findings. The viewing of pictures creates a memory representation that conceptualizes as information(Loftus et al., 1985, Whitley, 2013). Pictures, symbols and colors can be combined to create an infographic that tells a story (Krauss, 2012). The information represented as an image causes the brain to look at the information from more than one angle which often assists the brain in faster absorption of the information (Davison, 2009, Zhang, 2012).

By using Twitter, as a mobile learning technology, learning in the real world context is suddenly available. When receiving 140 characters as a SMS on a personal mobile device, thebrain automatically notes the physical context to where information was read and the mode ofdelivery. Adding more than one situational context can affect learning. Both physical location (Smith, 2013) and mode of delivery affect learning (Westera, 2011). Receiving a tweet on their mobile device has now provided an individual three memory cues for recalling the information:**where** one was when learning, **how** one received the information, and the **time** of day he/she received the information.

In addition to the contextual environment, mode of delivery and time of day, a tweet could ignite emotions that can later be associated with the information. The emotions could be aroused by the improper time in a life event when receiving the tweet, or an appropriate time of receiving a tweet (Cramp et al., 2012). Or, emotions could be stirred by the word content of the tweet. Fritsch and Kuchinke found enough difference in neuro-stimulation associated with known words to pseudowords to claim that recognized words had an emotional impact with individuals (Fritsch and Kuchinke, 2013). By tweeting information into the hands of the mobile device equipped student, there are multiple modes and models of learning that can be made available to aidin their learning.

3 PURPOSE

The purpose of this project is to learn if employing Twitter as a communication course requirement decreases the need to access the course CMS and results in a personalized multi-modal learning environment where studentsbecome more engaged with the course content and share more personal discoveries related to the course material in real-time.

4 PROJECT DESCRIPTION

This project uses Twitter (a mobile communications application primarily accessed via smart phones), Blackboard, (a web-based content management system (CMS)), and web videos, most likely hosted by a personal YouTube.com station or a video file uploaded to the school's webserver. Twitter will be used to supplement class communication hosted in the 'discussion' area of Blackboard, or through the 'announcement' area. All Twitter feeds are linked and stored in the Blackboard course for automatic archiving. These can be accessed at anytime through Blackboard.

Geographically diverse online graduate nursing informatics students will create a personalized Twitter account. With their Twitter account, they will interact with each other in real-time at anytime by sending tweets. A single class Twitter account has been established through GroupTweet.com. This software provides one Twitter account where multiple people can post and receive tweets. By using GroupTweet.com the students do not have to establish each classmate as a follower to their personal Twitter account. They only have to ensure that they follow the GroupTweet.com account established for the class. Establishing a GroupTweet.com account also permits the students and the instructor to maintain their personal Twitter account for personal tweets that are not to be shared with the class.

Students have access to assigned course readings and use the GroupTweet.com account to participate in class through Blackboard. Each week a different topic or aspect of information technology (IT) project management (PM) is introduced through lecture slides, a textbook, journal articles, blogs and newsfeeds. In the lecture material, the instructor provides an assigned task related to the weekly subject matter, or a question for the students to consider. Students create a personalized web video response for the question and then Tweet to the class that they have posted their web video response. Those receiving the tweet can then automatically open the video on their smartphone or their personal computer. Responses to the videos can be tweeted instantly upon review. This instantly connects the students at one time and documents the running dialogue that answers the posed question. It also eliminates the steps necessary to login to Blackboard to open the video, to simply add feedback.

During the 15-week course students are expected to tweet eight individual tweets to the class where they share a real-time life learning moment with the course content. This could simply be pushing a news article they discovered that related to one of the weekly topics, or it could be how they lived a PM moment in their life. This course requirement will bring 'real-life' into the virtual classroom and expose students to the practicalities of the course content. Within two of these eight 'life learning moments' they are to utilize the hashtag (#). This requirement has the potential to link previous tweets discussing the same topic throughout the term.

5 GOALS AND OBJECTIVES/RESEARCH QUESTIONS

Goals of this projectin no specific order are the following:

1) Utilize the Internet based Twitter communication software to establish a multi-modal student learning opportunity;
2) Foster the growth of a personal learning network (PLN) that originates from the classroom and becomes a sustaining PLN outside of the classroom;
3) Reduce the students' need to access the CMS through a desktop or laptop computer as all course announcements and student postings are pushed through Twitter in real-time.

6 METHODOLOGY AND TIMELINE

Pending IRB approval a convenience sample of 10 geographically diverse online graduate nursing informatics students will engage with global class learning techniques (Twitter, web-based videos, and a CMS) daily/weekly for 15 weeks beginning early 2014.

Evaluation of the global learning experience will be collected at three time points from students through a formal online qualitative and quantitative questionnaire. Questions will cover students initial interest in Twitter, if they had an existing Twitter account, self reported expertise with Twitter, self report of hashtag deployment, ability to set alerts on their mobile device and more. Any feedback offered without solicitation will also be reviewed. Themes and changes in qualitative information will be recorded and discussed by the research team. The quantitative information will be assessed by examining response rates as well as an examination of the distribution of a total score.

GroupTweet.com and Blackboard provide statistics on users. The feasibility of this project will be assessed by examining frequencies ofthe number of times the tools are accessed, the time of day the tools are accessed, the length of time the students spend with each tool, and reasons why the specified software tool was accessed. All student feedback will be utilized to modify or supplement the project as needed.

7 BENEFITS AND ANTICIPATED OUTCOMES

The outcomes of this project will inform future deployment of Twitter as a learning tool in the mobile classroom environment.

By requiring use of technologies, students adopt the technology, step away from their routine, and learn something new; even if it makes them uncomfortable. Using global class learning tools, e.g. Twitter, web-based videos, and a CMS, is certainly nothing new but how they are used and blended together pushes the students to bring 'real life' tools into their world of learning.

With the right kind of assignments and the use of Twitter, 'just-in-time' learning could occur daily – multiple times. This extends the classroom into real-life, making the didactic content practical. Students are exposed to real-life situations and real-time situations while connected to their personal learning network. Thus, there is greater likelihood that the students can apply what they learn and retain that learning for future use.

Connecting the students remotely and having just-in-time interaction might establish a personal learning network (PLN) that would continue to function as an established network after the course is completed. A PLN could become as popular as social support networks and professional networks.

The novelty of the hashtag could reveal subject matter themes throughout the term that could then be used to incorporate future course content. It can also be used to link conversations together and extend the conversation into the public world.

The 140 character restriction used by Twitter will require the students to become more succinct, if not more creative, in what they share; that is, they will have to get the point across quickly. By creating 'blasts of information', students are ultimately chunking what they have to say. This provides the link to Miller's 1956 Chunking Theory. The human brain can only retain a certain quantity of information in one short instance. Miller's findings demonstrated that the human brain is capable of retaining 7 numbers, plus or minus two, in its working memory. Receiving a blast of limited information within 140 characters might have the same retention properties as 7, plus or minus 2 numbers. If this is the case, the modes of delivering academic lecture material will change worldwide.

8 CONCLUSIONS

There is some evidence of Twitter being used in the classroom but not by many (Lin et al., 2013, Seaman and Tinti-Kane, 2013). Over 7,000 college and university faculty completed an online survey on using social media in the classroom. A little over 9% reported using Twitter professionally and 60% of those users were from the applied sciences discipline (Seaman and Tinti-Kane, 2013). The faculty claimed concern for personal privacy risks of students as their reason for not utilizing social media such as Twitter. To address this concern in advance, the author is utilizing services from a third party software company with a mechanism to protect personal tweets and tweet to a private group.

There is a plethora of research and knowledge about human learning styles, learning mechanisms, memory tricks and instructional methods. This paper presented numerous discoveries around contextual learning and memory capacity. Twitter presently appears to satisfy multiple learning styles and deliver a succinct message quickly.

Significant time was spent constructing solid scaffolding for methodically deploying and using Twitter into an online course managed through a CMS. The next phase of this study will use this model design and introduce Twitter as a learning mechanism in a more analytical setting; an introduction to statistics course. Findings from this additional phasewill be consulted to improve Twitter utilization for the faculty as well as the students.

REFERENCES

Cramp, A., Lamond, C., Coleyshaw, L. & Beck, S. 2012. Empowering or Disabling? Emotional Reactions to Assessment amongst Part-Time Adult Students. *Teaching in Higher Education,* 17**,** 509-521.

Davison, S. 2009. A Picture is Worth a Thousand Words. *Science and Children,* 46**,** 36-39.

Fritsch, N. & Kuchinke, l. 2013. Acquired Affective Associations Induce Emotion Effects in Word Recognition: An ERP Stud. *Brain and Language,* 124**,** 75-83.

Krauss, J. 2012. Inforgraphics: More than Words Can Say. *Learning & Leading with Technology,* 39**,** 10-14.

Lin, M.-F., Hoffman, e. & Borengasser, c. 2013. Is Social Media Too Social for Class? A Case Study of Twitter Use. *TechTrends,* 57**,** 39-45.

Loftus, G. R., Johnson, C. A. & Shimamura, A. P. 1985. How much is an icon worth? *Journal of Experimental Psychology: Human Perception and Performance,* 11**,** 1-13.

McKee, R. & Fryer, B. 2003. Storytelling that moves people. *Harvard Business Review,* 81**,** 51-55.

Miller, G. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review,* 63**,** 81-97.

SALOMAN, G. 1979. *Interaction of Media, Cognition, and Learning,* San Francisco, Jossey-Bass.

Seaman, J. & Tinti-Kane, H. 2013. Social Media for Teaching and Learning. Babson Survey Research Group.

Smith, P. R. 2013. Psychosocial Learning Environments and the Mediating Effect of Personal Meaning upon Satisfaction with Education. *Learning Environments Research,* 16**,** 259-280.

Westera, W. 2011. On the Changing Nature of Learning Context: Anticipating the Virtual Extensions of the World. *Educational Technology & Society,* 14**,** 201-212.

Whitley, C. T. 2013. A Picture Is Worth a Thousand Words: Applying Image-based Learning to Course Design. *Teaching Sociology,* 41**,** 188-198.

Zhang, Y. A. 2012. Developing Animated Cartoons for Economic Teaching. *Journal of University Teaching and Learning Practice,* 9.

FastTriaje: A Mobile System for Victim Classification in Emergency Situations

Alexandra Rivero-García, Candelaria Hernández-Goya, Iván Santos-González and Pino Caballero-Gil
Department of Computer Engineering, Universidad de La Laguna, Tenerife, Spain
{*alexandra.rivero.00, mchgoya, ivan.santos.99, pcaballe*}*@ull.edu.es*

Keywords: Triaje, FastTriaje, NFC, Android, Elliptic Curves, Lightweight Cryptography, Internet of Things, Zero Knowledge Proofs.

Abstract: The high penetration of communication technologies and smartphones may help in many complex scenarios. This work presents a system to perform victim diagnosis in emergency situation and/or natural disasters. The implementation include a web platform, a web service and a mobile application. The synergy among these three elements and different communication technologies such as NFC and Wi-Fi allows to classify potential casualties in a fast, useful and reliable way. Robust cryptographic methods are used to ensure the access only to legitimate users.

1 INTRODUCTION

The system proposed in this paper is called FastTriaje and it is based on the method START (Simple Triage and Rapid Treatment) (Iserson and Moskop, 2007). This method allows to obtain two essential aims in an emergency situation and/or a natural disaster: saving as many lives as possible and optimizing the available resources, both material and human ones (Raúl Sánchez Bermejo, 2011), (no,).

A traditional triage helps to decide on patient's attention priority by means of three actions: Inspection, Evaluation and Decision. With FastTriaje the disgnosis process is completely guided by the application. The system itself will indicate to the diagnostician the result of the evaluation and the decision to be carried out.

One of the system elements is a mobile application for Android devices. The application will be the diagnosticians tool to classify the victims and storing the information of each triage in an NFC tag that will be attached to the victim. The tag generated may be accessed at any time afterwards through the same application.

The second pillar in the system is a web platform whose main objectives are centralizing the triage information and easing user management.

The mobile application and the web platform interacts by means of a REST web service, through the transmission of different messages in JSON format.

Important security measures, based on lightweight cryptography and zero knowledge proofs have also been included improving reliability.

2 VICTIM CLASSIFICATION: TRIAJE SYSTEMS

The word TRIAGE is a term recently coined by the scientific community referring to patient classification. A commonly accepted definition follows: a simple, completed, objective and fast process to obtain an initial clinical assessment of people with the objective of evaluate their immediate survival capacities and prioritizing them according their severity.

An efficient method for casualty classification in a hostile situation is crucial. In order to achieve the classification, all triage systems distinguish two steps:

- First triage or simple triage. It is carried out in the hostile zone with fast methods like SHORT, START (Simple Triage and Rapid Treatment) or MRCC (Método Rápido de Clasificación en Catástrofes). Here, the diagnostician will not spend more than a minute, evaluating the survival capabilities of victims and sorting them depending on their seriousness.
- Second triage. It is made in a medical station or in a hospital. At this point, medical staff analyze each patient state: bruises, wounds and injuries. The Andorran Triage Model (MAT), Spanish Triage System (SET) and Manchester Emer-

gency Triage System (METS) are methods widely used during this stage.

Note that two types of triages are defined but they are not mutually exclusive, both should be carried out. The first triage gives important information about priorities for evacuation to hospitals where medical staff will do the second triage.

A more detailed description of the START method (Simple Triage and Rapid Treatment) has been included bellow because is the one used by FastTriaje, (see figure 1).

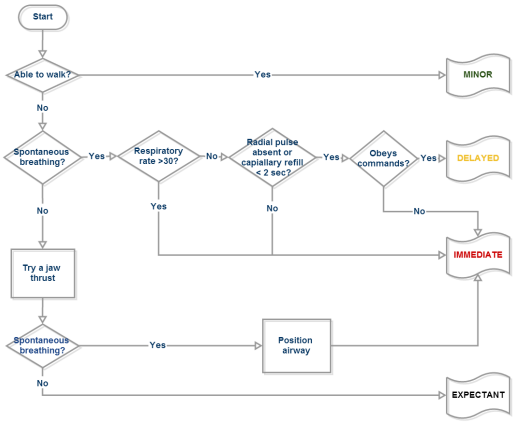


Figure 1: Simple Triage and Rapid Treatment Algorithm.

START method consists of answering a few simple questions to evaluate possible casualties. The answers are just “Yes” or “No” and depending on the answer given, the method will make another question or establish a classification. The questions are related to:

- mobility,
- breathing,
- perfusion, and
- state of mind.

START uses colored tags in the classification, each color stands for a different seriousness level:

- Black: dead or irrecoverable victims.
- Red: victims requiring immediate care.
- Yellow: victims requiring urgent care, but can wait for treatment from half an hour to one hour.
- Green: victims who are not seriously injured. They can wait for treatment more than an hour.

The traditional colored tags, are substituted by NFC tags. Next this technology and its security countermeasures are introduced.

3 NEAR FIELD COMMUNICATION TECHNOLOGY

Near Field Communications or NFC (nfc, b) is a short-range high frequency wireless communication technology. It may be defined as an extension of Radio Frequency Identification (RFID) since it combines the interface between tags and reader into a single device. It allows users to share content between digital devices in a peer to peer mode what makes it a promising paradigm in mobile technology.

In 2003 NFC was approved as ISO standard and nowadays is a technology used in many applications like car keys, ID cards, tickets, smartphones and so on. Contrary to other technologies as RFID, Bluetooth, ZigBee or Wi-Fi, NFC is not designed to send information continuously. Communication between devices requires that devices touch each other to exchange the information quickly and timely.

Some key features of NFC are:

- Its frequency range is 16.56MHz, this is a free frequency.
- Different communication rates are allowed: 106kb/s, 212kb/s or 424kb/s.
- High proximity between devices is needed: 10cm or less.
- It is an open technology based on standards.
- It may be considered a simple and reliable technology.
- Two communication modes are supported: active and passive mode. The first one allows communication between devices that can generate their own magnetic camp. The second is devoted to the interaction with NFC tags.

One of the main advantages of NFC when comparing it with other technologies is its inherent security. Due to its short communication range and the need of user involvement when an action is performed, NFC provides a higher security level.

A comparison among different technologies has ben included in Table I ((nfc, a)).

4 FastTriaje SYSTEM

FastTriaje main scenario is an emergency situation or natural disaster. Its goal is to speed up the classifications of victims. The modules that compose the whole system may be appreciated in figure 2: a mobile application and a web platform.

Table 1: Comparison of some short range technologies.

	Technology			
Feature	NFC	Bluetooth	RFID	ZigBee
Connection set	0.1s or less	6s	0.1s or less	30ms
Speed	424-848kbps	24Mbps (version 3.0)	424kbps	250kbps
Range	10cm	10m	3m	70m
Battery consumption	Low	High	Low	Low
Security	High	High	Vulnerable	Vulnerable
User experience	Touch	Need configuration	No configuration	No configuration



Figure 2: FastTriaje modules.

A mobile application is included. The main goal of this Android application is implementing the START triage method in mobile devices as a simple, usable and intuitive tool that facilitate the traditional triage process. All the triages generated are sent to a web platform, where every user has access to the information associated to the triages made by him/her. Additionally, the application allows to store each triage on an NFC tag to label victims.

Only legitimate users can exchange information with the web platform since a robust authentication protocol (see ZKP1 in next section) is used before transferring triages. All users who want to use the mobile application must be registered in advance on the web platform. User's privileges are set by the system administrator on this platform. Thus, it makes possible to distinguish between users with permissions to only read tags from those who may write on it, as well.

The web platform communicates with the mobile application by means of a web services. Before writing information on a NFC tag, the authorized user executes another authentication protocol (ZKP2) that associates the triage to be stored there with his/her credentials.

5 LIGHT WEIGHT CRYPTOGRAPHY AND ZERO KNOWLEDGE PROOFS

Nowadays mobile applications are designed for almost any field: business activities, transport management, social sharing and so on. In all of them guarantying information security is a must. Traditional solutions generally need specific infrastructure, transferring these solutions to the environment under study here is non-viable (Babar et al., 2010).

We have used lightweight cryptography in order to guarantee legitimate access to the information associated to the triage system. Concretely, we use Elliptic Curve Cryptography (ECC) (Hankerson et al., 2003) since:

- it provides with problems with higher computational complexity, and
- key length to achieve certain security level is shorter.

When using ECC there is certain notation related to the curves used that should be defined. This notation is included in table II.

$$E(F_p) = \{(x,y) \in F_p x F_p : y^2 = x^3 + ax + b\} \cup \{O\} \quad (1)$$

where

$$a, b \in F_p, 4a^3 + 27b^2 \neq 0$$

Table 2: Elliptic curve domain parameters in ZKPs.

Parameter	Description
p	prime number defining the field F_p
a, b	coefficients in the equation of elliptic curve E
P	a base point (a generator of a cyclic subgroup of $E(F_p)$)
m	order of P in $E(F_p)$

This election is justified by the restrictions on the computational and communication capabilities defined on the devices involved, smartphones mainly. The solutions developed are based on Zero Knowledge Proofs (ZKP) which stand out as a promising solution to the authentication problem in the Internet of Things (Martínez et al., 2009), (Ramzy and Arora, 2011), (Alpár et al., 2012).

These protocols allow a party A to convince another party B, about the knowledge of certain secret information without sending anything related to

it. These demonstrations may be extended to solve the authentication problem as it isincluded in the ISO standard 9798-5 dedicated to entity authentication.

ZKPs use three main elements:

- Witness (w): the prover selects a random item from a predefined set, keeping it secret. This value is called compromise (x). From it, the prover generates a value called witness, and then, sends it to the verifier.
- Challenge (e): In the second step, the verifier selects a random question that must be answered by the prover correctly, provided that he/she knows the secret information. This question is related to x and with the secret associated to the credentials of the prover.
- Answer (y): Finally the prover sends a response to the challenge and it is now when the verifier analyzes that response and checks if it is correct, in this case the authentication is accepted.

6 AUTHENTICATION METHODS IN FastTriaje

This section describes the authentication protocols implemented in FastTriaje to guarantee access control only to legitimate users. The two first protocols share certain features, for example both protocols are interactive (two parties, the prober A and the verifier B takes part) and the bootstrapping stage consists of defining an elliptic curve E and a base point P on it. Apart from that, the credentials associated to A are defined in the same way: the secret identification is an integer randomly selected in Zp, while the public one is a point on the curve E generated by multiplying the secret integer by the base point.

6.1 ZKP1 One-way Authentication: First Proposal

This protocol is devoted to authenticate mobile devices (A) against the web platform (B). A detailed description is included in Table III. It makes use of a hash function to define the challenge in each execution. This feature does not corresponds with the original definition of ZKPs but allows to reduce the number of iterations to just one in each authentication interaction.

Table 3: ZKP1.

Stages	Actions
Bootstrapping	p prime number E elliptic curve in Z_p $P \in E$
A's secret identification	$a \in Z_p$
A's public identification: PuidA	$a*P \in E$
Compromise: A's secret	$x \in_r {}^1Z_p$
Witness: $A \rightarrow B$	$w = x*P \in E$
Challenge: $A \leftarrow B$	$e = hash(P, a*P, x*P)$
Answer: $A \rightarrow B$	$y = x + a*e \in Z_p$
Verification: B checks	$y*P - e*PuidA = w$

6.2 ZKP2 One-way Authentication: Second Proposal

The protocol described here is used to authenticate a user (A) against an NFC tag (B) prior to storing a triage on it.

Table 4: ZKP2.

Stages	Actions
Bootstrapping	p prime number E elliptic curve in Z_p $P \in E$
A's secret identification	$a \in Z_p$
A's public identification: PuidA	$a*P \in E$
Compromise: A's secret	$\{x_1*P, x_2*P, \cdots, x_n*P\}$ $\in E$, with $x_i \in_r Zp$
Witness: $A \rightarrow B$	$w = hash(x_j*P + x_k*P)$, with $j,k \in_r \{1,2,\cdots,n\}$
Challenge: $A \leftarrow B$	$e \in_r Zp$
Answer: $A \rightarrow B$	$y = x_j + x_k - a*e \in Z_p$
Verification: B checks	$hash(y*P - e*PuidA)$ $= w$

However this protocol cannot be used directly on the NFC tags since they are totally passive, making it impossible for them to generate challenges. That is the reason why the the Fiat-Shamir Paradigm (Fiat and Shamir, 1987) has been used to transform the interactive ZKP2 into a non-interactive ZKP. The new protocol is described in Table 5.

Applying this paradigm requires the definition of the challenge by using a hash function. We have opted by using the new standard SHA3 (Bertoni et al.,).

Table 5: ZKP2 Non interactive.

Stages	Actions
Bootstrapping	p prime number E elliptic curve in Z_p $P \in E$
A's secret identification	$a \in Z_p$
A's public identification: PuidA	$a * P \in E$
Compromise: A's secret	$\{x_1 * P, x_2 * P, \cdots, x_n * P\}$ $\in E$, with $x_i \in_r Zp$
Witness: $A \rightarrow B$	$w = SHA3(x_j * P + x_k * P)$, with $j,k \in_r \{1,2,\cdots,n\}$
Challenge: $A \leftarrow B$	$e = SHA3(w) \in Z_p$
Answer: $A \rightarrow B$	$y = x_j + x_k - a * e \in Z_p$
Verification: B checks	$SHA3(y * P - e * PuidA)$ $= w$

7 CONCLUSIONS AND FUTURE WORK

A system that may improve logistics, the classification and attention of casualties in hostile situations such as natural disasters or accidents have been developed. The tool consists of a mobile application and a web service. The mobile application performs the triages and allows to store the results on NFC tags, that may be attached to victims. It is also possible to transfer them through a web service to a centralized web platform where the information may be processed according different user profiles.

Since the services provided are critical, security on data communication and reliability requirements have been taken into account.

There are still some points in the system than can be improved, such as:

- Adding statistical functionalities to the web platform.
- Integrating it with patient's clinical records.
- Extending the application to perform the second kind of triage.

ACKNOWLEDGEMENTS

This research has been partially supported by the Spanish Ministry of Economy and Competitiveness and Ministry of Science and Innovation under grants IPT-2012-0585-370000: DEPHISIT and TIN2011-25452: TUERI.

REFERENCES

NFC and Contactless Technologies. http://www.nfc-forum.org/aboutnfc/nfc_and_contactless/.

NFC Forum Official Page. http://www.nfcforum.org/home/.

Alpár, G., Batina, L., and Verdult, R. (2012). Using NFC phones for proving credentials. In *Proceedings of the 16th international GI/ITG conference on Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, MMB'12/DFT'12, pages 317--330, Berlin, Heidelberg. SpringerVerlag.

Babar, S., Mahalle, P., Stango, A., Prasad, N., and Prasad, R. (2010). Proposed security model and threat taxonomy for the internet of things (IoT). In *Recent Trends in Network Security and Applications*, volume 89 of *Communications in Computer and Information Science*, pages 420–429. Springer Berlin Heidelberg.

Bertoni, G., Daemen, J., Peeters, M., and Assche, G. V. The Keccak sponge function family . http://keccak.noekeon.org/papers.html.

Fiat, A. and Shamir, A. (1987). How to prove yourself: Practical solutions to identification and signature problems. pages 186--194. Springer-Verlag.

Hankerson, D., Menezes, A. J., and Vanstone, S. (2003). *Guide to Elliptic Curve Cryptography*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Iserson, K. V. and Moskop, J. C. (2007). Triage in medicine, part i: Concept, history, and types. *Annals of Emergency Medicine*, 49(3):275 – 281.

Martínez, S., Valls, M., Roig, C., Miret, J., and Giné, F. (2009). A Secure Elliptic Curve-Based RFID Protocol. *Journal of Computer Science and Technology*, 24(2):309–318.

no, J. A. M. C. El triaje. http://www.dit.upm.es/ jantonio/personal/cruzroja/.

Ramzy, I. and Arora, A. (2011). Using zero knowledge to share a little knowledge: Bootstrapping trust in device networks. In *Stabilization, Safety, and Security of Distributed Systems*, volume 6976 of *LNCS*, pages 371–385. Springer Berlin Heidelberg.

Raúl Sánchez Bermejo, Carmen Cortés Fadrique, B. R. F. (2011). El triaje en urgencias en los hospitales españoles. Unidad de Urgencias, Hospital General Nuestra Señora del Prado, Toledo, España.

SOCIETY, E-BUSINESS AND E-GOVERNMENT

FULL PAPERS

I Am off Then: Drivers of Travellers' Intentions to Book Trips Online
An Integrated Study on Technology Acceptance and Satisfaction

Maria Madlberger
Department of Business and Management, Webster University Vienna, Berchtoldgasse 1, 1220, Vienna, Austria
madlberger@webster.ac.at

Keywords: Electronic Commerce, Technology Acceptance Model, Satisfaction, Information Quality, System Quality, Online Booking Intention, Travel Agency, PLS Analysis.

Abstract: The tourism industry has undergone a substantial transformation since the emergence of electronic commerce. Especially travel agencies that are faced with growing online competition are increasingly dependent on achieving online sales. This study investigates antecedents of consumers' intention to book a trip online at a travel agency's website. The research draws on an integrated research model based on the technology acceptance model and customer satisfaction as introduced in the DeLone and McNeal model on information system success. An online survey among 292 consumers largely supports the hypothesized impact factors. Information quality serves as a significant object-based belief that influences satisfaction as an object-based attitude. In contrast, system quality has no impact on satisfaction. Satisfaction influences perceived usefulness, a key driver of online booking intention and perceived ease of use. The study provides several scholarly and managerial implications for the online distribution of tourism services.

1 INTRODUCTION

Online distribution of services constantly increased significantly the last years. One service sector that experienced substantial transformation is the tourism and travel industry. Travel agencies are a prominent example of e-commerce-induced disintermediation since their business has been increasingly replaced by direct online distribution of flights, hotel rooms, rental cars, organized tours, and other travel services. Worldwide, sales of traditional offline travel agencies are declining sharply (eMarketer, 2012).

However, travel agencies can offer online service provision themselves and thus defend their position in the distribution of tourism services. Many travel agencies operate websites that offer online search and online booking tools. The emergence of online mediators such as Expedia.com demonstrates that there is a significant demand for online platforms that offer a variety of tourism services which complement service providers' Web presences, such as airline or hotel websites. Furthermore, travel agencies frequently offer services that differ from single travel components. Especially when it comes to packaged tours or holiday arrangements travel agencies take the work of selecting the trip components out of the consumers' hands. Further, from a legal perspective, consumers who book services via a domestic travel agency may be in a better position in case of a dispute as their domestic law is applied instead of the law of the foreign country the service provider (e.g., a hotel) resides in.

Hence, despite the increased competition by tourism service providers, online intermediaries, and even consumers who plan their trips themselves, travel agencies offer a substantial value-added. On the other hand, more and more consumers expect online booking facilities and therefore travel agencies are increasingly dependent on attracting consumers who prefer to book online. The most obvious measure for a travel agency to gain online consumers is operating a website and/or mobile application that allows online booking of trips or single tourism services. For this purpose, deeper insights into drivers of online purchasing behavior are of crucial importance for travel agencies.

Although extensive research has been done on online consumer and purchasing behavior, findings on online booking behavior on travel agencies' websites are still limited. There is especially incomplete knowledge on antecedents that are based on external stimuli which can be controlled by companies when designing their websites. In order

to shed light on a comprehensive model on drivers of online booking intentions, we develop a research model that integrates two seminal theories in information systems (IS) and e-commerce research: the technology acceptance model (TAM; Davis, 1989) and the notion of satisfaction grounded in the DeLone and McNeal information system success model (DeLone and McLean, 1992; DeLone and McLean, 2003). In doing so, our study follows the theoretical approach developed by Wixom and Todd (2005) and further extended by Xu et al. (2013). In line with the suggestion by Ajzen and Fishbein (1980), we contend that online booking intention is the result of a chain of impacts that starts with external stimuli that lead to object-based beliefs (information quality and system quality) and object-based attitude (satisfaction with the website) which itself influences behavioral beliefs and behavioral attitudes or behavioral intentions (Wixom and Todd, 2005; Xu et al., 2013). The proposed structural model is tested with data collected from an online survey among 292 consumers. The partial least squares (PLS) analysis results confirm most of the proposed hypotheses.

The study contributes to research by stressing the relevance, particularly of information quality, and satisfaction for booking intention on travel agencies' websites. It further demonstrates the relevance of the TAM in the online journey booking context. This research also offers important managerial implications by showing key design issues of travel agencies' websites to maximize online booking intention. The paper is organized as follows: The next section presents the theoretical background, particularly on the TAM and the role of service quality and satisfaction in the online tourism sector. The following section three presents the research model and the hypotheses development. The subsequent section four shows the research methodology of the survey. Section five presents the results which are discussed in section six. The paper proposes research and managerial implications and closes with further research directions.

2 THEORETICAL BACKGROUND

2.1 TAM and Satisfaction in Online User Behavior

Like other IS-related issues on user behavior, online consumer behavior, especially usage and intention to use has been extensively and successfully investigated through the theoretical lens of TAM (Davis, 1989; Davis et al., 1989). TAM is grounded in the theory of reasoned action (TRA; Ajzen and Fishbein, 1980; Fishbein and Ajzen, 1975) and its extension, the theory of planned behavior (TPB; Ajzen, 1985; Ajzen, 1989). TPB extends the notion of TRA that beliefs influence attitudes which themselves have an impact on behavioral intentions. The TAM applies this theory to the context of IS usage where perceived usefulness (PU) and perceived ease of use (PEOU) are considered main drivers of behavioral attitude, intention, and behavior. PU, the main impact factor, is defined as "the degree to which a person believes that a particular system would enhance his job or performance" (Davis, 1989, p. 320) while PEOU is a secondary impact factor and denotes "the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989, p. 320). Further, PEOU has an indirect positive impact on system usage via PU (Davis, 1989).

In IS and e-commerce research TAM was constantly extended and further developed. An early extension was suggested by Davis (1993) who considers system design features an external stimulus that precedes PU and PEOU. In e-commerce research TAM was applied and extended to predict website use and online shopping behavior in numerous studies (for example Ahn et al., 2004; Ha and Stoel, 2009; Klopping and McKinney, 2004; Lin and Lu, 2000; Liu and Forsythe, 2010; McCloskey, 2004; McCloskey, 2006; Moon and Kim, 2000; Shih, 2004; Wang and Benbasat, 2005). Significant modifications of TAM in the context of e-commerce and WWW usage were made by Gefen and Straub (2000), Lederer et al. (2000), Teo et al. (1999), and Klopping and McKinney (2004) who successfully simplify the TAM model by eliminating behavioral attitude and confirming a direct impact of PU and PEOU on behavioral intention. Another significant contribution was made by researchers who drew on Davis' (1993) extension of TAM by external stimuli. Several researchers adopted the components of the DeLone and McLean IS success model (DeLone and McLean, 1992) or the updated DeLone and McLean IS success model (DeLone and McLean, 2003) including information quality, system quality, and service quality (see Brown and Jayakody, 2008 for a review). Examples of studies using information quality, system quality, and service quality as extensions of TAM in e-commerce are Shih (2004) and Ahn et al. (2004). Another construct of the DeLone and McLean model,

satisfaction, was compared with TAM (Wang, 2008) or successfully integrated into TAM-based research on system use as in the seminal studies by Wixom and Todd (2005) and Xu et al. (2013). These studies consider object-based beliefs in the form of information quality and system quality (Wixom and Todd, 2005) as well as service quality (Xu et al., 2013) as antecedents of object-based attitude, expressed as satisfaction with these quality dimensions which further influence PU and PEOU.

2.2 Consumer Behavior in Online Tourism

The tourism industry is characterized by a number of particularities that differentiate it from other domains, such as durable and non-durable consumer goods. Like many services, tourism services are perishable, non-storable, differing in quality, and largely influenced by the consumers themselves. Particularly vacation trips can be highly complex products consisting of a number of single components that are difficult to assess by consumers. Since the prices of travel products are usually high in relation to the income, consumers travel rather rarely and thus are often not very experienced (Järveläinen, 2007). Travel services are further a typical experience good, that is, consumers can evaluate it only after consumption. As a result, consumers' expectations when purchasing travel services are different from expectations when purchasing physical consumer goods. For example, prior to booking, consumers need information on hotel locations, flight times, or costs and distances of rental cars or public transportation (Petre et al., 2006). Thus an e-commerce site that offers tourism services needs to provide necessary information that allows a sufficient assessment of the travel products prior to booking a trip.

The tourism industry is one of the service sectors that was most influenced by e-commerce (Kim et al., 2011; Lin et al., 2009). Especially the number of traditional offline travel agencies declined with increased competition from service providers such as airlines or hotels, but also online intermediaries like Expedia or Travelocity. Worldwide total travel sales reached estimated $ 962 billion in 2012, with $ 374 billion or 38.9 percent share achieved online. In the United States this share is 51.5 percent, in Europe it amounts to 45.1 percent.

Classical travel agencies were constantly losing market shares. This development can be observed throughout Western Europe where in 2012 only a fifth of trips was booked via travel agencies while in 2008 it was every third trip (Hotelmarketing.com, 2013). Data from the United States indicate that this trend does not necessarily affect travel agencies as an institution, but the offline distribution channel. In 2011, 91 percent of active travelers booked their trips online whereas the traditional offline travel agency was employed in 9 percent of the cases. 62 percent of travelers used online travel agencies, followed by branded supplier sites (e.g., airlines) with 46 percent, meta search sites (14 percent) and collective buying as well as private sales sites (5 percent each; Mashable, 2013). This evidence stresses the relevance of the Internet as a distribution channel that is increasingly becoming a "must" for travel agencies. Whereas these figures evoke the impression that travel agencies are in a good position in the online business, these numbers also show that particularly classical travel agencies nowadays are faced with a much larger and more complex competition than they were in the declining offline distribution. For example, travel agencies are becoming increasingly challenged by online platforms such as Expedia.com in the domain of packaged tours (Dooley, 2009), a very important business segment of travel agencies.

Online consumer behavior in the tourism industry was subject of several empirical studies. Particular focus is put on satisfaction and website quality. For example, Jeong et al. (2003) investigate website quality and information satisfaction as impact factors of behavioral intention. Law and Ngai (2005) consider the usability of a website a key dimension of website quality. Tsang et al. (2010) analyze dimensions of service quality of travel agencies and distinguish website functionality, information quality and content, fulfillment and responsiveness, safety and security, appearance and presentation, and customer relationship and confirm their significant impact on satisfaction. The relations between process and outcome quality, satisfaction, and behavioral intention are investigated by Chen and Kao (2010). Similar findings are achieved by Hsu et al. (2012) who add perceived playfulness and perceived flow as variables. A comprehensive model including website quality, satisfaction, trust, attitude, and behavioral intention is provided by Wen (2012).

Also research based on TAM was carried out in several empirical studies. TAM was augmented with various factors, such as trust and perceived risks (Nunkoo and Ramkissoon, 2013), task ambiguity, product complexity, and consumer experience (Järveläinen, 2007), or more comprehensive sets of variables like website content issues, previous visits, and accessibility (Kaplanidou and Vogt, 2006).

Chang et al. (2012) integrate perceived website quality into the TAM by considering perceived website quality gain and loss (quality appraisal) as antecedents of PU and PEOU. Quality appraisal is derived from the IS success model by DeLone and McLean (2003) and is composed of information, system, and service quality. Park et al. (2007) investigate the impact of several dimensions of website quality including PEOU on willingness to use. Ryan and Rao (2008) test the TAM in the tourism context in its original form, supplemented by system security. The above-mentioned studies highlight the appropriateness of the TAM as well as the DeLone and McNeal IS success model for explaining drivers of website use in the context of traveling and tourism.

3 RESEARCH MODEL AND HYPOTHESES DEVELOPMENT

The proposed research model is grounded in TAM and its extension by Wixom and Todd (2005). In line with the original DeLone and McLean IS success model (DeLone and McLean, 1992), Seddon (1997), and Wixom and Todd (2005), we consider two dimensions of website quality perception, that is information quality and system quality, as we are solely focusing on the intention to use a travel agency's website and not the full offered spectrum of a travel agency's services.

Figure 1 summarizes the research model.

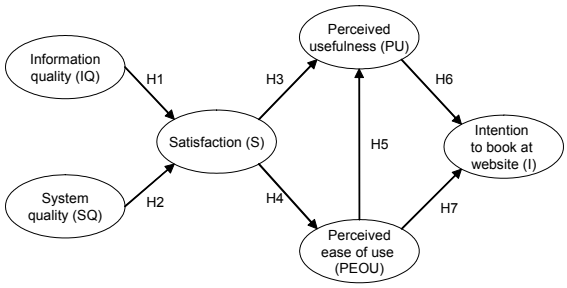


Figure 1: Research model.

TAM proposes that behavioral attitudes and behavior are driven by the behavioral beliefs PU and PEOU. As Wixom and Todd (2005) and Xu et al. (2013) argue, these behavioral beliefs are themselves influenced by object-based beliefs and attitudes. In particular behavioral beliefs are influenced by object-based attitudes, that is, satisfaction with the system, which itself is impacted by object-based beliefs, that is, quality perceptions. This approach is different from research that suggests quality perceptions as direct antecedents of PU and PEOU (Ahn et al., 2004; Shih, 2004).

Among the quality perception variables, information quality denotes the perceived quality of the offered content. Information quality should assist consumers in their shopping process by facilitating the comparison of products, increasing enjoyment, and improving decision-making. In the context of e-commerce, information quality is related to the content of the website (Ahn et al., 2004). Characteristics that represent information quality are timeliness, completeness, and accuracy of information (DeLone and McLean, 2003). We thus contend that websites that fulfill these criteria are perceived as well-performing on information quality. System quality describes the technical performance as well as the design of the web site as an information system and thus refers to the engineering perspective (Ahn et al., 2004). Perceived system quality includes attributes such as reliability, flexibility, or availability (DeLone and McLean, 2003; Wixom and Todd, 2005) as well as functionality that may influence an online shopping process (Shih, 2004).

There is ample e-commerce literature that demonstrates the relevance of information quality and system quality for perceptions of website quality (for example Aladwani and Palvia, 2002; Kim and Stoel, 2002; Lin and Lu, 2000; Liu and Arnett, 2000). While Wixom and Todd (2005) differentiate between information and system satisfaction and Xu et al. (2013) further introduce service satisfaction, we consider a more parsimonious model including general satisfaction with the website as proposed by DeLone and McLean (2003). We thus hypothesize:

H1: Perceived information quality of the website positively influences satisfaction with the website.

H2: Perceived system quality of the website positively influences satisfaction with the website.

Satisfaction with a website is an attitude toward an object (Ajzen and Fishbein, 1980; Wixom and Todd, 2005) that can be understood as an external variable that influences behavioral beliefs, namely PU and PEOU (Xu et al., 2013). Since PU is understood as the perceived degree to which a system increases the performance of the undertaken task (Davis, 1989), Wixom and Todd (2005) conclude that higher information satisfaction will be positively associated with PU. Likewise, as PEOU describes the perception that using a system does not require much effort (Davis, 1989), satisfaction with the system is expected to impact PEOU (Wixom and

Todd, 2005). In our more parsimonious approach, we expect that the object-based attitude satisfaction will positively influence the behavioral beliefs simultaneously. We therefore hypothesize:

H3: Satisfaction with the website positively influences perceived usefulness of the website.

H4: Satisfaction with the website positively influences perceived ease of use of the website.

The remaining variables are directly derived from the rich literature on TAM which consistently investigates and empirically confirms the impact of the behavioral beliefs PU and PEOU on behavioral attitude (for example Ahn et al., 2004; Ha and Stoel, 2009; Lin and Lu, 2000; McCloskey, 2004; McCloskey, 2006; Shih, 2004) or, in a more parsimonious way, on behavioral intention (Gefen and Straub, 2000; Klopping and McKinney, 2004; Lederer et al., 2000; Teo et al., 1999). In line with this literature as well as Wixom and Todd (2005) and Xu et al. (2013), we hypothesize:

H5: Perceived ease of use of a website positively influences perceived usefulness of the website.

H6: Perceived usefulness of a website positively influences the intention to use the website for booking.

H7: Perceived ease of use of a website positively influences the intention to use the website for booking.

4 RESEARCH METHODOLOGY

4.1 Instrument Development

The empirical test of the research model took place by means of a quantitative consumer survey. In order to achieve comparable results, respondents were presented a website of a travel agency that was subject to evaluation based on the applied variables. The measurement of items was done on the basis of elaborated scales from IS literature (Ahn et al., 2004; Devaraj et al., 2002; McCloskey, 2004; Shang et al., 2005; Shih, 2004). All items on consumers' beliefs and behavioral intentions were related to the presented website. Where necessary, the formulation was adapted to that context (e.g., "booking on the Website"). PU was measured with items adapted from Shang et al. (2005), PEOU items were adapted from Shih (2004) and McCloskey (2004), the items on information and system quality are based on Shih (2004) and Ahn et al. (2004). Finally, satisfaction and intention to use were adapted from Devaraj et al. (2002). Since the used items were all developed in English language, they were translated into German and back-translated by a native speaker. A pretest among eight students was made. According to these refinements a few wording modifications were done. All items were measured with a 5-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree).

4.2 Sample

The research design comprised a quantitative online survey. In order to attract consumers who are interested in traveling, the online survey was announced at several German-speaking online forums. 324 questionnaires were completed, among which 32 contained incomplete answers so that 292 questionnaires were used for further analysis. The gender distribution in the sample is 54.3 percent males and 45.6 percent females. The average age is 33 years; 26.2 percent are younger than 25 years, 38.3 percent are between 25 and 34 years, 19.5 percent are between 35 and 44 years, and 16.1 percent are 45 years or older. A filter question ensured that respondents travel at least once per year.

5 RESULTS

5.1 Measurement Model

The research model was tested by means of a Partial Least Squares (PLS) analysis. The used analysis software was SmartPLS (Ringle et al., 2005). The test of the measurement model consists of analyzing the consistency (Cronbach's Alpha), the convergent and the discriminant validity. Table 1 shows the Cronbach's Alpha and AVE values of the variables.

Table 1: Realiability measures of variables.

Variable	Number of items	Cronbach's Alpha	AVE
Information quality (IQ)	3	0.70	0.61
System quality (SQ)	4	0.84	0.67
Perceived usefulness (PU)	4	0.69	0.51
Perceived ease of use (PEOU)	3	0.78	0.70
Satisfaction (S)	3	0.86	0.78
Intention to use (I)	3	0.90	0.83

The Cronbach's Alpha values are, with one exception, higher than the recommended value of 0.7 (Nunnally, 1978). For PU, Cronbach's Alpha is 0.69 and therefore very close to the recommended value. Convergent validity is satisfactory if the

average variance extracted (AVE) is higher than 0.5 (Fornell and Larcker, 1981). This condition is met for all variables. Table 2 displays the numbers concerning discriminant validity.

Table 2: Correlation matrix.

	IQ	SQ	PU	PEOU	S	I
IQ	*0.78*					
SQ	0.39	*0.82*				
PU	0.34	0.05	*0.71*			
PEOU	0.21	0.21	0.43	*0.84*		
S	0.35	0.18	0.52	0.48	*0.88*	
I	0.38	0.06	0.66	0.42	0.66	*0.91*

In Table 2, the correlations of the variables are shown. The numbers on the diagonal in italics are the square roots of the AVE. For adequate discriminant validity, these values should exceed the interconstruct correlations. This condition is met for all constructs. Further, the loadings of the individual items on the corresponding variables are well above the recommended value of 0.5 for appropriate discriminant validity. They range between 0.66 and 0.94. Thus overall the measurement model is highly satisfactory.

5.2 Hypotheses Test

The test of the structural model comprises the path coefficients, the R-square values of dependent variables as well as the p-values. The latter were obtained by bootstrapping with 100 cases and 1,000 samples. The R-square values are the following: 0.12 for satisfaction, 0.31 for PU, 0.23 for PEOU, and 0.46 for booking intention. Table 3 shows the results of the PLS analysis along with the p-values of the path coefficients.

Table 3: PLS analysis results.

Hypothesized impact	Path coefficient	p-value
Information quality -> satisfaction (H1)	0.333	***
System quality -> satisfaction (H2)	0.045	n.s.
Satisfaction -> PU (H3)	0.409	***
Satisfaction -> PEOU (H4)	0.476	***
PEOU -> PU (H5)	0.235	*
PU -> Intention to book (H6)	0.593	***
PEOU -> Intention to book (H7)	0.163	n.s.
p-values: *** < 0.001, * < 0.05, n.s. not significant		

As the results of the PLS analysis show, five out of seven hypotheses are supported by data. Information quality shows a high positive impact on satisfaction with the travel agency website, supporting H1. In contrast, system quality shows a path coefficient close to zero, thus H2 is rejected. The impacts of satisfaction on PU and PEOU are both strong and highly significant (0.409 and 0.476, respectively), thus supporting H3 and H4. The impact of PEOU on PU is smaller (0.235), but still significant at the 5 percent level. Finally, intention to book at the travel agency's website is largely influenced by PU (0.593), supporting H6. The impact of PEOU is weak (0.163) and although there is a tendency of significance (less than ten percent), H7 is rejected.

6 DISCUSSION

6.1 Discussion of Results

The analyzed research model demonstrates the appropriateness of an integrated approach based on the DeLone and McNeal IS success model as well as TAM. The seminal work by Wixom and Todd (2005) could overall be confirmed in a service-oriented setting, that is, online booking at travel agencies' websites. The hypotheses tests, however, require a differentiated view on the analyzed variables. Perceived information quality shows a positive and highly significant impact on satisfaction, thus confirming our hypothesis.

Unlike expected, perceived system quality shows no significant effect. There are several possible reasons that system quality does not impact satisfaction. First, with increasing maturity and technical reliability of websites, the relevance of this factor may be decreasing over time. System quality is mainly caused by an advanced technical basis which can be expected to improve with growing IS sophistication of travel agencies. Second, system quality may play a minor role especially in the travel booking context where consumers may put a larger emphasis on the complex travel products and therefore on the information quality. Third, a possible explanation may lie in the parsimonious approach of this research that considered overall satisfaction rather than differentiation between information and system satisfaction. The results of this study, however, are consistent with a study on Web service quality differences between online travel agencies and online service providers that identifies information content as the most important dimension of Web service quality for online travel agencies (Kim and Lee, 2004).

The impact of satisfaction on the TAM-based constructs PU and PEOU could clearly be confirmed. With path coefficients higher than 0.40

(PU) and 0.47 (PEOU), these impacts are clearly shown. The results confirm the theoretical assumption based on Ajzen and Fishbein (1980) that object-based attitudes (satisfaction) have an impact on behavioral beliefs (PU and PEOU). The strongest impact throughout the model is found between PU and intention to book at the website (path coefficient of 0.593). This finding is highly consistent with previous research on TAM that identified PU as the primary impact factor on behavioral attitude and behavioral intention. On the other hand, the impact of PEOU on behavioral intention is not significant despite a slight tendency. This result is also consistent with previous research on TAM that shows mixed empirical evidence on the impact of PEOU (Lin and Lu, 2000; Moon and Kim, 2000; Venkatesh, 2000; Venkatesh and Davis, 2000). Further, a study on website quality and behavioral beliefs on websites of different tourism companies showed that PEOU is of less importance for travel agencies' websites while it is a primary factor for online service providers (Kim and Lee, 2004). Finally, as proposed in many TAM-based studies, the impact of PEOU on PU is significant, too.

6.2 Research Implications

The study at hand provides several theoretical and managerial implications. From the scholarly viewpoint, the results show the relevance of two seminal theories in IS research – the DeLone and McNeal IS success model and TAM – in the online tourism sector. It confirms the relevance of integrating both theories for a better understanding of online usage intention and thus supports Ajzen and Fishbein's (1980) notion of an impact chain starting with external stimuli of the website that drive object-based beliefs, object-based attitudes, behavioral beliefs, and behavioral intention. It particularly shows that the chain of impacts on intention to book online starts with information quality through satisfaction, PEOU, and PU, the latter being the main direct antecedent. Hence, both satisfaction and PEOU have a mediated impact on booking intention.

6.3 Managerial Implications

This research has important managerial implications. In the light of an increased dominance of online travel booking and a disintermediation threat of classical, offline travel agencies the drivers of online booking behavior at travel agencies' websites become increasingly essential. The significant role of information quality that ultimately impacts the intention to book at the agency's website stresses the importance of a careful design of the offered content on the website. Perceived information quality can be controlled by a website operator. Information quality can be enhanced by a simple enrichment of the website content, for example by offering multimedia contents including animations and videos. A growing number of hotels offer virtual tours or 360 degree views of hotel facilities to give a realistic impression of the property. Besides operator-generated contents, travel agencies should further make use of user-generated contents by providing space for user recommendations, reviews, and numerical ratings. Such consumer reviews are a core part of online traveling platforms like Expedia.com or Tripadvisor.com (Park and Allen, 2013) and can further enhance the information quality of travel agency websites.

Although system quality does not show a significant impact on satisfaction, attention should be paid on a high degree of system availability, security, and reliability. Of further importance is the key role of PU that has a much higher weight compared with PEOU. Hence a travel agency website should offer not only all necessary information and functionalities that enable online booking, but support and facilitate the all transaction phases of a trip booking. This includes a comprehensive after-sales service, for example by offering the provision of online feedback by customers after the trip. Since travel agencies usually provide additional information in printed catalogs and the agency bureaus, they should pay large attention on avoiding outdated facts or information that is inconsistent with information provided on the website.

Recently, the use of mobile devices for information search and booking trips is increasing sharply along with a switch of users between devices. Today, a typical "journey" across the devices may start with information search on the smart phone or tablet and finish with the booking process via the laptop or PC (Marketingcharts.com, 2013). Travel agencies have to account for this development and must offer a seamless and integrated information provision and booking process across these access devices without interruption. In the light of these developments, the impact of system quality on satisfaction should be revisited.

In a further step, travel agencies could even start learning from successful e-commerce retailers like Amazon.com by offering recommendation systems

based on consumers' past search and booking behavior. Particularly vacation trips vary considerably in terms of destinations, trip duration, travel time, activities, hotel categories etc. that can serve as useful criteria for customized recommendations. Last but not least, experience goods, such as vacation travel is often subject to extensive word-of-mouth on travel experiences among consumers. Hence social media like Facebook or Twitter should be considered important elements of the online marketing mix of travel agencies (Sotiriadis and Zyl, 2013).

7 CONCLUSIONS

This study offers a parsimonious model on drivers of intentions to book at travel agencies' websites based on Wixom and Todd's (2005) integration of TAM with the DeLone and McNeal IS success model. The results stress the opportunities for travel agencies to influence online purchasing behavior positively by offering information quality that satisfies users. The main contribution of this study to research lies in the investigation of drivers of usage intention at travel agencies' e-commerce sites and thus confirming Wixom and Todd's (2005) model in this context. From the managerial perspective, the study provides a theory-based framework on important website design issues that are critical for satisfaction, PU, PEOU, and usage intention which are important prerequisites of online booking behavior. Although the study was done in the context of a rather complex product, the findings can be transferred to other domains that involve experience goods or complex shopping goods, too. Business models in practice show that useful key features on e-commerce sites, such as multimedia information or user-generated content are used in an ever-increasing number of industries and product categories.

Although the results largely support the assumptions of the research model, there are several limitations of this study. First, we consciously decided to consider information and system quality as object-based beliefs. Some studies on website quality in e-commerce also consider service quality which is worth being investigated as a further independent variable. Second, we did not differentiate between different kinds of trips and journeys which may result in different impact strengths of the antecedents. For example, a packaged far-distance tour that takes three weeks consists of a series of service components and thus requires more information than booking a flight that can be described with few and structured pieces of information. Also differences between private and business trips may occur. Finally, socio-demographic factors and personal traits (e.g., traveling behavior or destination preferences) may have an impact on the overall proposed antecedents.

Further research should consider emerging e-commerce developments, especially the role of mobile devices and social media as well as the presence of online consumer reviews. Moreover, an analysis of different players in the online tourism sector (online service providers, electronic intermediaries etc.) should be compared with travel agencies to further increase the understanding of drivers of booking trips online.

REFERENCES

Ahn, T., Ryu, S., and Han, I., 2004. The impact of the online and offline features on the user acceptance of Internet shopping malls, *Electronic Commerce Research & Applications,* 3, 4, 405-420.

Ajzen, I., 1985. From intentions to actions: a theory of planned behavior, in *Action Control: From Cognition to Behavior,* J. Kuhl and J. Beckmann (eds.), Springer: New York, NY, 11-39.

Ajzen, I., 1989. Attitude structure and behavior, in *Attitude Structure and Function,* A. R. Pratkanis, S. J. Breckler and A. G. Greenwald (eds.), Lawrence Erlbaum: Hillsdale, NJ, 241-274.

Ajzen, I., and Fishbein, M., 1980. *Understanding Attitudes and Predicting Social Behavior*, Prentice-Hall: Englewood Cliffs, NJ.

Aladwani, A. M., and Palvia, P. C., 2002. Developing and validating an instrument for measuring user-perceived Web quality, *Information & Management,* 39, 6, 467-476.

Brown, I., and Jayakody, R., 2008. B2C e-commerce success: a test and validation of a revised conceptual model, *The Electronic Journal Information Systems Evaluation* 11, 3, 167-184.

Chang, K.-C., Chen, M.-C., Hsu, C.-L., and Kuo, N.-T., 2012. Integrating loss aversion into a technology acceptance model to assess the relationship between a website quality and website user's behavioural intentions, *Total Quality Management,* 23, 8, 913-930.

Chen, C.-F., and Kao, Y.-L., 2010. Relationships between process quality, outcome quality, satisfaction, and behavioural intentions for online travel agencies - evidence from Taiwan, *Service Industries Journal,* 30, 12, 2081-2092.

Davis, F. D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly,* 13, 3, 319-339.

Davis, F. D., 1993. User acceptance of information technology: system characteristics, user perceptions

and behavioral impacts, *International Journal of Man-Machine Studies,* 38, 3, 475-487.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R., 1989. User acceptance of computer technology: a comparison of two theoretical models, *Management Science,* 35, 8, 982-1003.

DeLone, W. H., and McLean, E. R., 1992. Information systems success: the quest for the dependent variable, *Information Systems Research,* 3, 1, 60-95.

DeLone, W. H., and McLean, E. R., 2003. The DeLone and McLean Model of information systems success: a ten-year update, *Journal of Management Information Systems,* 19, 4, 9-30.

Devaraj, S., Fan, M., and Kohli, R., 2002. Antecedents of b2c channel satisfaction and preference: validating e-commerce metrics, *Information Systems Research,* 13, 3, 316-333.

Dooley, G., 2009. Analysis: can travel agents compete against online tour sellers?, http://www.travelagentcentral.com/home-based/tours-tour-operators/analysis-can-travel-agents-compete-against-online-tour-sellers-15638.

eMarketer, 2012. Online travel sales explode in Latin America, http://www.eMarketer.com/Article/Online-Travel-Sales-Explode-Latin-America/1009493.

Fishbein, M., and Ajzen, I., 1975. *Belief, Attitude, Intention, and Behavior*, Addison-Wesley: Reading, MA.

Fornell, C., and Larcker, D. F., 1981. Evaluating structural equation models with unobservable variables and measurement error, *Journal of Marketing Research,* 18, 1, 39-50.

Gefen, D., and Straub, D., 2000. The relative importance of perceived ease-of-use in IS adoption: a study of e-commerce adoption, *Journal of the Association for Information Systems,* 1, 8, 1-21.

Ha, S., and Stoel, L., 2009. Consumer e-shopping acceptance: antecedents in a technology acceptance model, *Journal of Business Research,* 62, 5, 565-571.

Hotelmarketing.com, 2013. Travel agencies versus the Internet: global booking trends, http://www.hotelmarketing.com/index.php/content/article/travel_agencies_versus_the_internet_global_booking_trends/.

Hsu, C.-L., Chang, K.-C., and Chen, M.-C., 2012. The impact of website quality on customer satisfaction and purchase intention: perceived playfulness and perceived flow as mediators, *Information Systems & e-Business Management,* 10, 4, 549-570.

Järveläinen, J., 2007. Online purchase intentions: an empirical testing of a multiple-theory model, *Journal of Organizational Computing and Electronic Commerce,* 17, 1, 53-74.

Jeong, M., Oh, H., and Gregoire, M., 2003. Conceptualizing website quality and its consequences in the lodging industry, *International Journal of Hospitality Management,* 22, 2, 161-175.

Kaplanidou, K., and Vogt, C., 2006. A structural analysis of destination travel intentions as a function of Web site features, *Journal of Travel Research,* 45, 2, 204-216.

Kim, M. J., Chung, N., and Lee, C. K., 2011. The effect of perceived trust on electronic commerce: shopping for online tourism products and services in South Korea, *Tourism Management,* 32, 2, 256-265.

Kim, S., and Stoel, L., 2002. Dimensional hierarchy of retail Web quality, *Information & Management,* 41, 5, 619-633.

Kim, W. G., and Lee, H. Y., 2004. Comparison of Web service quality between online travel agencies and online travel suppliers, *Journal of Travel & Tourism Marketing,* 17, 2/3, 105-116.

Klopping, I. M., and McKinney, E., 2004. Extending the technology acceptance model and the task-technology fit model to consumer e-commerce, *Information Technology, Learning & Performance Journal,* 22, 1, 35-48.

Law, R., and Ngai, C., 2005. Usability of travel websites: a case study of the perceptions of Hong Kong travelers, *Journal of Hospitality and Leisure Marketing,* 13, 2, 19-31.

Lederer, A. L., Maupin, D. J., Sena, M. P., and Zhuang, Y., 2000. The technology acceptance model and the World Wide Web, *Decision Support Systems,* 29, 3, 269-282.

Lin, J. C. C., and Lu, H., 2000. Towards an understanding of the behavioural intention to use a web site, *International Journal of Information Management,* 20, 3, 197-208.

Lin, P. J., Jones, E., and Westwood, S., 2009. Perceived risk and risk relievers in online travel purchase intentions, *Journal of Hospitality Marketing and Management,* 18, 8, 782-810.

Liu, C., and Arnett, K. P., 2000. Exploring the factors associated with Web success in the context of electronic commerce, *Information & Management,* 38, 1, 23-33.

Liu, C., and Forsythe, S., 2010. Sustaining online shopping: moderating role of online shopping motives, *Journal of Internet Commerce,* 9, 2, 83-103.

Marketingcharts.com, 2013. Travel research and booking seen a multi-device process, http://www.marketingcharts.com/wp/interactive/travel-research-and-booking-seen-a-multi-device- rocess-29809/?utm_campaign=newsletter&utm_source=mc&utm_medium=textlink.

Mashable, 2013. The evolution of online travel - infographic, http://mashable.com/2012/02/21/online-travel-infographic/.

McCloskey, D. W., 2004. Evaluating electronic commerce acceptance with the technology acceptance model, *Journal of Computer Information Systems,* 44, 2, 49-57.

McCloskey, D. W., 2006. The importance of ease of use, usefulness, and trust to online consumers: an examination of the technology acceptance model with older consumers, *Journal of Organizational & End User Computing,* 18, 3, 47-65.

Moon, J. W., and Kim, Y. G., 2000. Extending the TAM for a world-wide-web context, *Information & Management,* 38, 4, 217-230.

Nunkoo, R., and Ramkissoon, H., 2013. Travelers' e-purchase intent of tourism products and services, *Journal of Hospitality Marketing and Management,* 22, 5, 505-529.

Nunnally, J. C., 1978. *Psychometric Theory*, McGraw-Hill: New York, NY.

Park, Y. A., Gretzel, U., and Sirakaya-Turk, E., 2007. Measuring website quality for online travel agencies, *Journal of Travel and Tourism Marketing,* 23, 1, 15-30.

Park, S.-Y., and Allen, J. P., 2013. Responding to online reviews: problem solving and engagement in hotels, *Cornell Hospitality Quarterly,* 54, 1, 64-73.

Petre, M., Minocha, S., and Roberts, D., 2006. Usability beyond the website: an empirically-grounded e-commerce evaluation instrument for the total customer experience, *Behaviour and Information Technology,* 25, 2, 189-203.

Ringle, C. M., Wende, S., and Will, A., 2005. SmartPLS 2.0, SmartPLS: Hamburg, Germany, www.smartpls.de.

Ryan, C., and Rao, U., 2008. Holiday users of the Internet — ease of use, functionality and novelty, *International Journal of Tourism Resarch,* 10, 4, 329-339.

Seddon, P. B., 1997. A respecification and extension of the DeLone and McLean model of IS success, *Information Systems Research,* 8, 3, 240-253.

Shang, R.-A., Chen, Y. C., and Shen, L., 2005. Extrinsic versus intrinsic motivations for consumers to shop on-line, *Information & Management,* 42, 3, 401-431.

Shih, H.-P., 2004. An empirical study on predicting user acceptance of e-shopping on the Web, *Information & Management,* 41, 3, 351-368.

Sotiriadis, M., and Zyl, C., 2013. Electronic word-of-mouth and online reviews in tourism services: the use of Twitter by tourists, *Electronic Commerce Research,* 13, 1, 103-124.

Teo, T. S. H., Lim, V. K. G., and Lai, R. Y. C., 1999. Intrinsic and extrinsic motivation in Internet usage, *Omega,* 27, 1, 25-37.

Tsang, N. K. F., Lai, M. T. H., and Law, R., 2010. Measuring e-service quality for online travel agencies, *Journal of Travel & Tourism Marketing,* 27, 3, 306-323.

Venkatesh, V., 2000. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model, *Information Systems Research,* 11, 4, 342-365.

Venkatesh, V., and Davis, F. D., 2000. A theoretical extension of the technology acceptance model: four longitudinal field studies, *Management Science,* 46, 2, 186-204.

Wang, W., and Benbasat, I., 2005. Trust in and adoption of online recommendation agents, *Journal of the Association for Information Systems,* 6, 3, 72-101.

Wang, Y.-S., 2008. Assessing e-commerce systems success: a respecification and validation of the DeLone and McLean model of IS success, *Information Systems Journal,* 18, 5, 529-557.

Wen, I., 2012. An empirical study of an online travel purchase intention model, *Journal of Travel & Tourism Marketing,* 29, 1, 18-39.

Wixom, B. H., and Todd, P. A., 2005. A theoretical integration of user satisfaction and technology acceptance, *Information Systems Research,* 16, 1, 85-102.

Xu, J. D., Benbasat, I., and Cenfetelli, R. T., 2013. Integrating service quality with system and information quality: an empirical test in the e-service context, *MIS Quarterly,* 37, 3, 777-A779.

TweetPos: A Tool to Study the Geographic Evolution of Twitter Topics

Maarten Wijnants, Adam Blazejczak, Peter Quax and Wim Lamotte
Hasselt University - tUL - iMinds, Expertise Centre for Digital Media,
Wetenschapspark 2, 3590 Diepenbeek, Belgium
firstname.lastname@uhasselt.be, adam.blazejczak@student.uhasselt.be

Keywords: Twitter, Social Networking Sites (SNSs), Social Media, TweetPos, Geographic Trends, Investigative Tool.

Abstract: Popular Social Networking Sites (SNSs) like Twitter and Facebook are evolving into crowd-sourced, inter-disciplinary sensor systems that "monitor" a wide spectrum of (physical) properties and topics. This paper introduces *TweetPos*, a web service that is intended to facilitate the analytical study of geographic tendencies in Twitter data feeds. To oblige the human cognitive features, the TweetPos tool maximally relies on visual data structures like heatmaps and charts to represent the geo-spatial sources of tweets. The tool compiles data bodies that grant insight in both past and present tweet posting behavior, incorporates an animation engine to highlight temporal trends, and leverages layered visualization techniques so that multiple topics can be offset against each other, all from a geographic perspective. Via the presentation of two representative use cases, we comprehensively demonstrate TweetPos' data mining and analytical features and we illustrate the (geo-spatial) intelligence they can amount to. Thanks to a generic implementation, the TweetPos service is not geared towards a specific target audience but instead is sufficiently versatile to be valuable for a vast and varied collection of consumer profiles like social scientists and market analysts.

1 INTRODUCTION

Social Networking Sites (SNSs) were conceived as a means to virtually connect users and to offer them an intuitive forum to ubiquitously contribute and disseminate information in real time. As their number of subscribers rose over time, so did the amount of content that is managed by SNSs. As a result, they nowadays host a wealth of user-generated data that is highly heterogeneous in nature.

Over the years, SNSs have also evolved functionality-wise. While many such services were purely text-based upon their inception, they nowadays typically grant users the option to attach multimedia items like pictures and video clips to their contributions. Another feature that has become nearly commonplace in the SNS landscape, is geotagging (i.e., attaching geographic coordinates as metadata to messages). It is apparent that such novel facilities embellish the core SNS content and further extend its value.

Given their popularity and broad adoption, it is becoming evermore valid to regard SNSs as real-life, real-time and crowd-sourced sensor systems that generate valuable, highly heterogeneous data feeds (see, for example, (Sakaki et al., 2010)). Stated differently, popular present-day SNSs are rapidly transforming into representative data providers. By intelligently exploiting the data feeds that can be accumulated from them, innovative and value-added services can be conceived. In addition, mining and analyzing the information that is shared by end-users through social media can lead to valuable insights and knowledge. Possible application domains include consumer behavior modeling, consumer profiling, intelligent recommendation systems, and population sentiment assessment. Extracting such kinds of intelligence from SNSs however typically requires external tools, as profound mining and analysis mechanisms by default are lacking from their feature set.

In this paper, we tend to Twitter, the authoritative microblogging platform in the western world, and we focus on investigating the data that is hosted by this SNS from a geo-spatial perspective. In particular, we introduce the web-based *TweetPos* tool, a convenient means to display and study the geographic origin of tweets, and to uncover the geographical evolution of the popularity of tweet topics. A hybrid visualization method encompassing both heatmap- and chart-based data representation allows for thorough analysis and mining with regard to the geo-spatial distribution of tweeted material over time. The TweetPos web service affords keyword-based topic selection and includes a layering system that allows for easy comparison of the geographical trends of multiple subjects.

Furthermore, our tool is able to compile data sets that integrate a representative sample of tweets from the recent past with present-day tweet messages that are captured in real time, in order to grant insight in both historical and current tweet posting behavior. Finally, the accumulated data collections can be aggregated and studied on either a per-day or per-hour basis to provide some degree of analytical granularity. We argue that, combined, these features offer all necessary measures to perform significant research about the geographical sources of Twitter data. We will back this claim by presenting the results of two prototypical analyses that illustrate the versatility, effectiveness and comprehensiveness of the proposed instrument. At the same time, the provided demonstrations serve as prove of the extensive applicability of TweetPos: courtesy of its generic methodology, it may one way or another cater to the demands of a variety of human investigators, including social researchers, marketeers, analysts and journalists.

A primordial aspect of the TweetPos solution is its emphasis on providing graphical representations of the crawled Twitter data. Contrary to computers, the typical human mind does not excel at handling large quantities of raw data. On the other hand, our cognitive features make us more adept than computers at interpreting visual data structures (Pinto et al., 2010) like heatmaps and charts, which are exactly the output modalities that are supported by our platform. The TweetPos tool is hence intended to offer human operators an adequate graphical workspace that allows them to readily and conveniently assess geo-spatial trends in social media contributions.

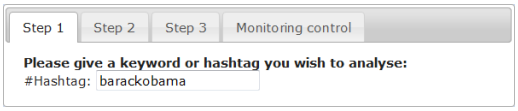
The remainder of this article is organized as follows. Section 2 presents an overview of the functional features of the TweetPos web service. Next, Section 3 handles the architectural design and implementation of the tool. We then evaluate our work in Section 4 by discussing some representative examples of investigations into the geographical evolution of recently trending Twitter themes that have been produced with the proposed tool. Section 5 briefly reviews related work on the analysis and mining of information that has been shared via social networks, and at the same time highlights our scientific contributions. Finally, we draw our conclusions and suggest potential future research directions in Section 6.

2 TweetPos

The TweetPos instrument is implemented as a web service that is accessible via a standard web browser. Screenshots of the tool's input widgets are bundled in Figure 1. As these images illustrate, keywords or so-called Twitter *hashtags* are the service's essential ingress parameters. Based on the specified topic of interest, the tool will compile a corpus of tweets that deal with this subject. This corpus will encompass a representative sample of historical messages as well as a completely accurate set of current and future tweets on the topic at hand. The user is hereby granted the option to apply geographical filtering by limiting the tweet compilation to either Europe or North America, if so desired (see Figure 1(b)). An identical filtering option is included in the input pane that controls the visualization of the accumulated data (see Figure 1(c)). Finally, a number of standard HTML input elements allow for controlling the temporal constraints and the animation of the result set. In particular, via two HTML sliders and a checkbox, users can enforce the discrete time interval with which (the timestamps of) gathered tweets need to comply for them to be included in the output. Two fixed levels of granularity are supported for the specification of the temporal constraints, which cause TweetPos to aggregate filtered tweets per hour and per day, respectively. An animation engine that utilizes either hourly or daily increments allows for the animated, video-like presentation of the tweet data set and as such might yield valuable insights into the geo-spatial trends that are exhibited by tweet topics over time.

On the output front, the principal GUI element consists of a topographic map that scaffolds heatmap-based visualization of the geo-spatial provenances of filtered Twitter messages. Stated differently, this output component displays the intensity, from a geographic point of view, of tweets that encompass the specified input keyword. Besides a map, two additional output widgets are included in the tool. The first is a line chart that visualizes the quantitative volume of the compiled tweet archive, aggregated either on a per-hour or a per-day basis, while the second enumerates the textual contents of the collected tweets. Figure 2 illustrates the TweetPos output interface.

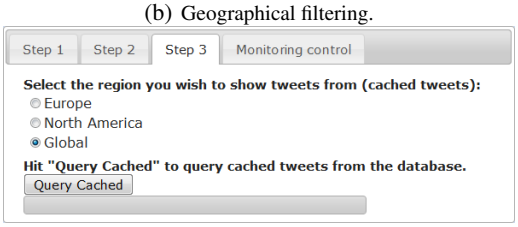
An important feature of TweetPos is its keyword layering functionality. The tool allows multiple keyword filters to be active simultaneously, by conceptually associating (the results of) each concurrent hashtag search with an individual layer. Figures 2(a) and 2(b) for instance illustrate a setup in which two queries are involved. Layers are rendered on top of the topographic map as uniquely colored overlays, whose visualization can be independently toggled on and off. Analogously, distinct tweet volumes are plotted in the line graph for each currently deployed keyword filter. A layer can be eliminated from the visualization process via the legend that is incorporated in



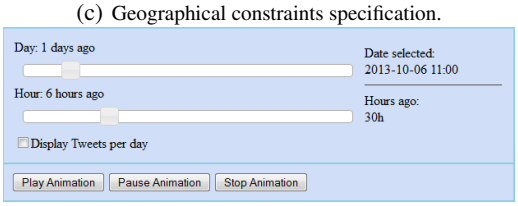
(a) Topic selection.



(b) Geographical filtering.



(c) Geographical constraints specification.



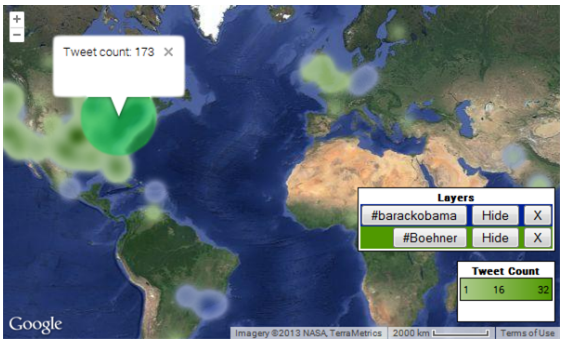
(d) Temporal constraints specification and animation control.

Figure 1: TweetPos input GUI.

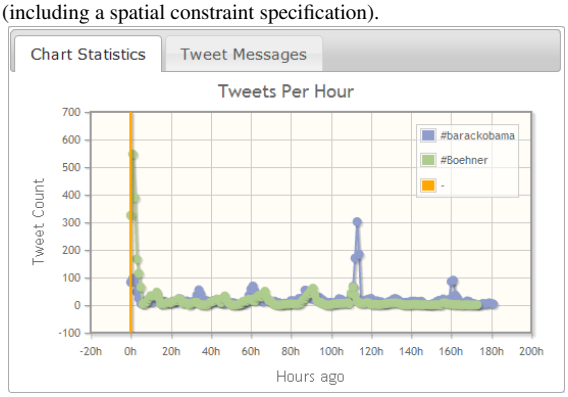
the geographic map. The layering system provides a powerful means to investigate (the geo-spatial evolution of) multiple subjects concurrently, to offset them against each other, to reveal potential correlations between them, and so on.

Apart from temporal filtering parameters, the TweetPos service also supports the specification of spatial constraints. This type of constraint is deployed by clicking on the topographic map, which causes a circular area to be drawn around the selected location (see Figure 2(a)). The map's zoom level and the stretch of the marked geographical region have been designed to be inversely proportional properties, which implies that the spatial extent of the highlighted area is controllable by zooming the map in and out. In effect, installing a spatial constraint under a relatively high zoom level will result in the selection of a relatively tight geographical region, while the opposite holds true when the map is heavily zoomed out.

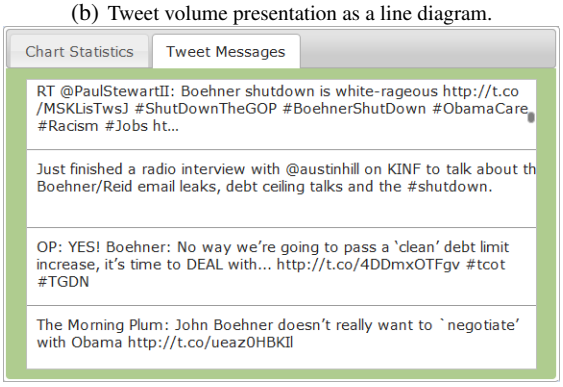
All output components are dynamic, in the sense that their content is updated on-the-fly when the user modifies one or more input parameters. Obviously this applies to the keywords or hashtags that are searched for. In particular, initiating a new search operation causes an additional layer to be introduced in



(a) Heatmap-based output of tweet locations on a topographic map (including a spatial constraint specification).



(b) Tweet volume presentation as a line diagram.



(c) Textual contents of the filtered tweets.

Figure 2: TweetPos output GUI.

both the 2D map and the line chart. Responding to less profound input settings however also occurs in real time. For example, exploiting the HTML sliders to modify the time constraints causes the map, the line chart as well as the list of tweet message to be updated instantaneously. The map will be adjusted to draw the geographic intensity that applied at the specified timestamp, the volume plot will be updated so that it correctly marks the currently selected time, and the textual list will only display tweet messages that satisfy the installed temporal restrictions. Analogous actions are dynamically undertaken in reaction to the definition of a spatial constraint. More precisely, the volume plot and textual message list only

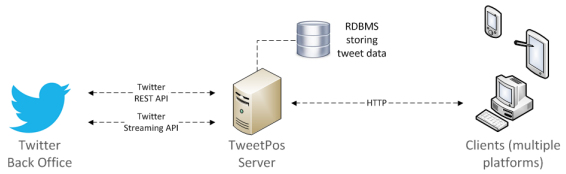


Figure 3: High-level system architecture.

reckon with tweets that originated from the designated spatial area, if any. This feature allows human operators to zoom in on certain geographic regions and to perform fine-grained, localized analyses. As a final example of the dynamism of the output GUI, switching between layers via the legend in the topographic map causes the contents of the textual tweet enumeration widget to be updated so that it only displays those messages that apply to the keyword that corresponds with the currently selected layer.

3 IMPLEMENTATION

The TweetPos implementation is completely web-compliant. HTML and CSS are used for rendering the GUI and for handling page layout and style, while all programmatic logic is scripted in PHP and JavaScript (at server and client side, respectively).

Our motivations for realizing the TweetPos application as a web service are manifold. First of all, selecting the web as deployment platform acknowledges the pervasiveness of the Internet in modern society. At the same time, it renders the TweetPos functionality available on all environments and devices that support widespread and standardized web technologies, which maximizes the portability of our implementation. Finally, numerous utility libraries and supportive tools exist for the web, which we have gladly leveraged to expedite the development process.

3.1 Architectural Design

A schematic overview of TweetPos' architectural setup is given in Figure 3. TweetPos adopts a client/server network topology. The back-end HTTP server forms the heart of the system; it interfaces with Twitter, implements the data filtering and compilation, hosts a relational database (RDBMS) for data persistence purposes, and responds to incoming HTTP requests. The client on the other hand is very lightweight, as its responsibilities are limited to user interfacing and data visualization. As such, the server (and the RDBMS which it encapsulates) forms a level of abstraction in the TweetPos system architecture between respectively the external information source (i.e., Twitter) and the client-side presentation of the disclosed data.

3.2 Twitter Data Collection

Twitter provides multiple HTTP-based APIs to enable third-party software developers to interface with the platform and to build socially-inspired applications. The TweetPos tool exploits two of these APIs in order to harvest both historical and up-to-date (public) Twitter data. First of all, the Twitter Search API (which is embedded in the Twitter REST API as of version 1.1) is leveraged to compose a non-exhaustive yet representative sample of tweets from the past 7 days that dealt with a particular subject. The quantitative incompleteness is intrinsic to Twitter and represents a deliberate strategy in the platform's design (Twitter Developers, 2013). In effect, the Search API has been designed for relevance and not completeness, which implies that it is not intended to deliver a rigorous index of past tweets. The second Twitter interface that fuels TweetPos' data collection procedure is a low-latency gateway to the global stream of tweets, called the Streaming API. This particular API allows developers to set up a long-lived HTTP connection to the Twitter back office, over which tweets from that moment on will then be streamed incrementally. In combination with extensive filtering and querying mechanisms, applications in this way obtain near-real-time and exhaustive access to exactly the type of tweets they are interested in. To facilitate the interaction with the Twitter Streaming API, the TweetPos tool integrates the 140dev Streaming API framework (140dev, 2013).

For the sake of comprehensiveness, we will now describe the complete set of actions and operations that constitute TweetPos' data ingestion pipeline. When a user initiates a new data collection process by transmitting a keyword-based query to the TweetPos server, the latter will spawn a total of seven PHP daemons. Each of these background processes utilize the Twitter Search API to jointly compile a pool of relevant historical tweets that were contributed during the past week (i.e., one process per day). At the same time, the back-end server manages a (PHP-based) daemon that permanently monitors the Twitter Streaming API. As an end-point is only allowed to set up a single connection to the Streaming API, this background process runs a cumulative filter to guarantee that all present and future tweets that satisfy one of the currently active queries are captured. In contrast to the Search API daemons, which have a finite execution time and are query-specific, the Streaming API process runs indefinitely and is shared by queries.

A dedicated widget in the client-side GUI empowers users to stop the real-time monitoring of a particular topic (which is enforced by updating the cumulative filter of the Streaming API daemon).

3.3 Data Storage and Processing

Fetched tweets are persisted at server side in a MySQL database. To streamline the integration of the 140dev framework in the TweetPos tool, we have opted to integrally adopt its cache architecture and accompanying database schema. The caching mechanism of the 140dev framework applies a two-step approach. An aggregation step continuously filters JSON-encoded tweet data (including the actual message and all sorts of metadata) from the Twitter Streaming API and inserts the resulting data directly into a designated caching table in the back-end database. In effect, this task is fulfilled by the Streaming API daemon that was mentioned in Section 3.2. Simultaneously, an independent background process successively pulls single raw JSON items from this table, parses and conveniently formats the composing entities of the corresponding tweets (i.e., the textual message itself, the encapsulated hashtags and mentions, etcetera), and distributes the outcome across dedicated database tables. By isolating the aggregation from the parsing of relevant tweets, real-time and lossless data ingestion is guaranteed (the Twitter Streaming API might yield tremendous quantities of data, whose sheer volume might prohibit on-the-fly parsing and processing).

Besides leveraging the 140dev caching methodology and database schema for the Streaming API context of the TweetPos tool, we have decided to extend their application to the Twitter Search API component of our implementation. This entails that historical tweets that are harvested by the Search API daemons are just as well cached in raw JSON format and then parsed by the same process that also handles Streaming API contributions. The beneficial implications of this design are that it yields a clean software architecture, ensures uniform treatment of tweets originating from heterogeneous sources, and enables the elimination of data duplication in an integrated manner (i.e., without requiring an exogenous control loop).

Once the data collection procedure for a particular keyword-based query has been initiated, all client requests that are related to this query are handled at server side by means of pure RDBMS interactions. As an example, the execution of adequate SQL statements suffices for the server to be able to forward an up-to-date overview of Twitter data pertaining to the queried topic to the client.

3.4 Geocoding

As the TweetPos tool is chiefly concerned with the geo-spatial provenance of tweets, it is clear that geographic metadata plays a primordial role in its operation. To be more precise, geographic coordinates are needed in order to pinpoint a tweet on a topographic map. Some Twitter users include these coordinates directly in their posts (e.g., users with smartphones with built-in GPS receivers), yet the majority only inserts a descriptive representation of the involved location (e.g., in the form of a textual address), or even leave out all geographic references altogether.

TweetPos' data accumulation procedure is agnostic of the presence of geo-spatial metadata in tweets. Stated differently, tweets that lack any trace of geographical metadata are not filtered out by either the Streaming API or Search API data compiler. Tweets holding exact geographic footprints are directly cached, as they can be readily localized on a map. In case the tweet only incorporates a descriptive geo-spatial reference, the data processing daemon described in Section 3.3 will invoke the Google Geocoding API (Google Developers, 2013b) to translate the description into geographic coordinates prior to database insertion. Finally, although non-localized contributions are not exploitable in the current implementation, they are still recorded in the database "as is" for the sake of completeness (i.e., they may hold some value in future extensions of the tool).

3.5 Visualization

All visualization and GUI interaction operations are performed at client side by means of HTML and JavaScript.

3.5.1 Heatmap-based Geolocation Clustering

The topographic output map has been implemented by means of the JavaScript variant of the Google Maps API (Google Developers, 2013a). Tweets are positioned on this map on the basis of the geographic location from which they were posted. Instead of marking (the location of) individual tweets on the map, a heatmap-based design has been adopted. Heatmaps are a general-purpose data visualization technique in which the intensity of data points is plotted in relative comparison to the absolute maximum value of the data set. Typically, data point intensity is indicated by means of a color coding scheme. Compared to mashups of discrete markers (which might easily clutter the map in the case of voluminous data sets), heatmaps hold the perceptual advantage

that, without sacrificing much detail, they are naturally surveyable and interpretable. The Google Maps JavaScript API has built-in support for heatmap rendering.

3.5.2 Line Graph

While the heatmap at a glance provides users with an impression of the spatial characteristics of a particular Twitter topic, it fails to communicate exact quantitative figures concerning the tweet volume. To counter this deficiency, the TweetPos tool includes a line graph visualization that discretely plots, either per hour or per day, the number of tweets that address the queried subject(s). As such, it visualizes a precise overview of the temporal evolution of the popularity of themes (expressed in tweet quantity). The line diagram is implemented via jqPlot, a plotting and charting plug-in for the jQuery JavaScript framework (http://www.jqplot.com/). The data values that compose the graph are interactive in the sense that they can be clicked to leap the date selection sliders (see Figure 1(d)) to the corresponding timestamp.

3.5.3 Tweet Message Enumeration

The TweetPos tool is also able to output the textual contents of filtered tweets. This output method has been realized by means of the MegaList jQuery plug-in (http://triceam.github.io/MegaList/). Like the other output widgets, it is adaptive in the sense that it dynamically adjusts its contents to imposed spatiotemporal constraints. This widget is intended to provide users insight into the context in which the queried topic is referenced. As such, it allows for accurate, context-aware classification of tweets based on the messages they carry. For instance, a tweet about a certain incident might plead for or, conversely, against it; by inspecting the textual context, the stance of the tweet publisher becomes apparent.

4 EVALUATION

This section serves to showcase the capabilities of the TweetPos instrument by presenting two representative examples of (geo-spatial) analyses of Twitter content that have been produced with it. The first test case is intended to rigorously demonstrate TweetPos' overall practicalities and to generally exemplify the data mining options which the tool scaffolds, while the second example focuses on TweetPos' layering functionality and the analytical features it entails. Space limitations force us to be brief in our discussion, and prevent us from including additional demonstrations.

4.1 2014 FIFA World Cup Qualifiers

The final two qualifier matches for next year's soccer World Cup were played on October 11th and 15th, 2013, respectively. We have exploited the TweetPos service to investigate the (geographic) resonance of these matches on Twitter, specifically for Belgium's national soccer team (which are nicknamed the "Red Devils" or "Rode Duivels" in Dutch). We issued a TweetPos data collection request for the `RodeDuivels` hashtag on October 13th and kept this query active until October 19th. Figure 4 shows the geographic distribution of the tweets that were gathered worldwide in the one hour interval immediately succeeding the end of the two matches, as well as a chart-based representation of the tweet quantity that was harvested during the entire course of the experiment (aggregated per hour). As the query was initiated on October 13th, all tweet data in the result set that precedes this date was acquired via the Search API, while tweets with an older timestamp were filtered from the Streaming API.

Analysis of the experimental results yields four notable observations. First and foremost, the output graph reveals two obvious peaks in tweet volume. These local maxima coincide nicely with the Red Devils' schedule of play. As such, this test case corroborates Twitter's capacity to act as a user-driven distributed sensor system that is able to identify real-world events (see also Section 5). As the data collection procedure was started in between the two matches, this capacity applies to both the Search API (for events from the recent past) and Streaming API (for current and future events). Secondly, tweets dealing with the match on October 11th appear to have originated practically exclusively from Belgium and its surrounding countries. In contrast, tweets about the second game exhibit a quasi worldwide distribution, yet again with a strong concentration in Western Europe. As the first set of tweets was ingested via the Twitter Search API, this outcome can likely be attributed to the operational principles of this interface (recall from Section 3.2 that the Search API aims for relevance, not comprehensiveness). Thirdly, although their volume is rather marginal, tweets embodying the `RodeDuivels` keyword were found to also emerge from non-Dutch speaking countries like the USA, Spain and Turkey (see the rightmost topographic map in Figure 4). After inspecting the textual contents of these contributions (by means of the tweet message enumeration widget described in Section 3.5.3), it became clear that these types of tweets can roughly be classified into two categories:

- tweets written in Dutch by Belgian citizens (tem-

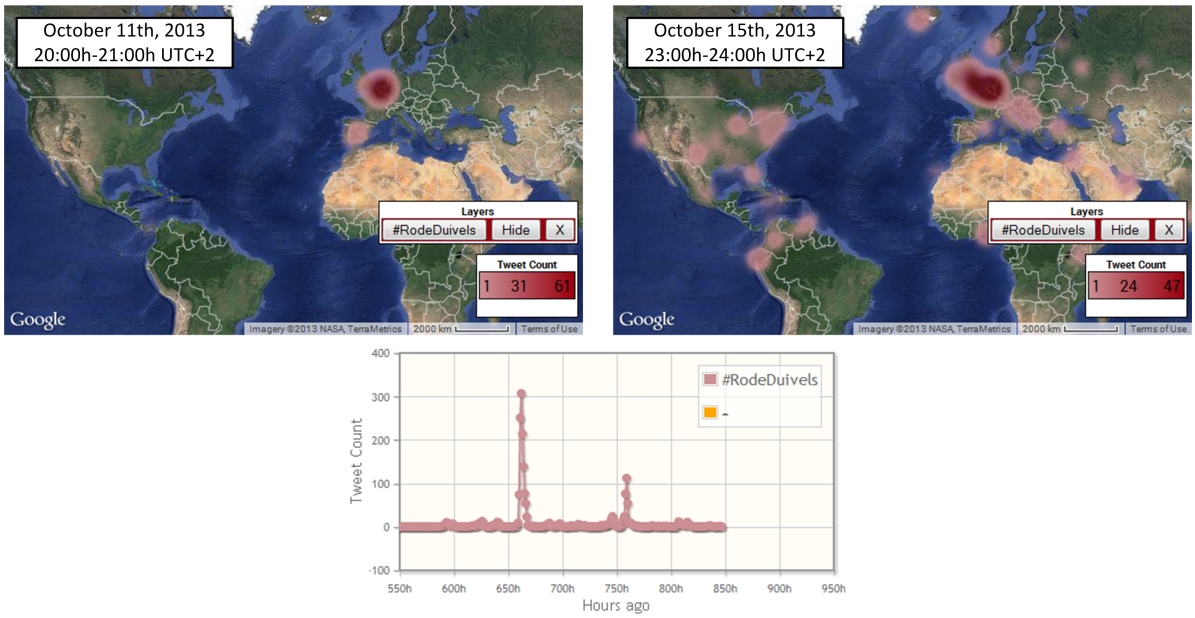


Figure 4: Results of the 2014 FIFA World Cup qualifiers experiment.

porarily) living abroad; e.g., "Come on #RodeDuivels, I am rooting for you from my hotel room in Barcelona!" (English translation)

- retweets by the local population of English messages that include the (Dutch) `RodeDuivels` hashtag; often, the original messages were posted by Dutch natives who wanted to reach an internal audience; e.g., "Belgium versus Wales qualifier starting in 15 minutes #RodeDuivels #RedDevils #belwal #wc2014"

The fourth and final observation pertains to location-driven personalization of the tweeted contents. For example, a tweet by Toby Alderweireld (a Belgian soccer player who plays for Atletico Madrid in Spain), written in English and communicating Belgium's qualification for next year's World Cup, was actively retweeted by his followers in Spain and amounted to the majority of `RodeDuivels` tweets that originated from that country. A single Spanish Atletico Madrid fan mentioned not only Toby Alderweireld but also his Belgian teammate Thibaut Courtois in his tweet: "Well done to #Atleti's @thibautcourtois & @AlderweireldTob and their #RodeDuivels teammates. We'll see you in Brazil at #wc2014".

4.2 Game Console Comparison

The market of (next-gen) gaming consoles is (for the time being) dominated by Sony, Microsoft and Nintendo with their PlayStation 4, Xbox One and Wii U hardware, respectively. In this second test case, the TweetPos tool was put to use to compare the attention these three consoles receive on the Twitter network, and to uncover geographic dissimilarities between their respective popularity, if any. Therefore, between November 1st and November 16th, 2013, the `ps4`, `xboxone` and `WiiU` keywords were tracked with TweetPos. An impression of the resulting data set is given in Figure 5. This figure visualizes the geospatial intensities of the three hashtags on the launch day of the PlayStation 4 in the USA (i.e., on November 15th between 07:00h and 08:00h UTC-5), as well as per-hour aggregated overviews of the volumetric magnitudes of the collected data sets.

These experimental results validate that TweetPos succeeds in layering multiple heatmaps, each associated with an independent query, on top of a single topographic map. The same holds true for the tweet volume plotting functionality of the line chart. Notice however from the topmost row of images in Figure 5 that keyword visualizations might quickly conceal one another in multi-layer scenarios, which in turn is likely to impair analytical efficiency. Courtesy of TweetPos' ability to on-the-fly switch the rendering of individual layers on and off, it nonetheless remains feasible to interactively compare and interpret (the geographic provenance of) tweets in multi-query studies. In effect, the images in the bottom three rows in Figure 5 communicate exactly the same information as the ones in the upper row, yet in an itemized fashion.

In-depth analysis of the composed data body falls beyond the scope of this article. Instead, we will point out two illustrative insights that we were able to ex-

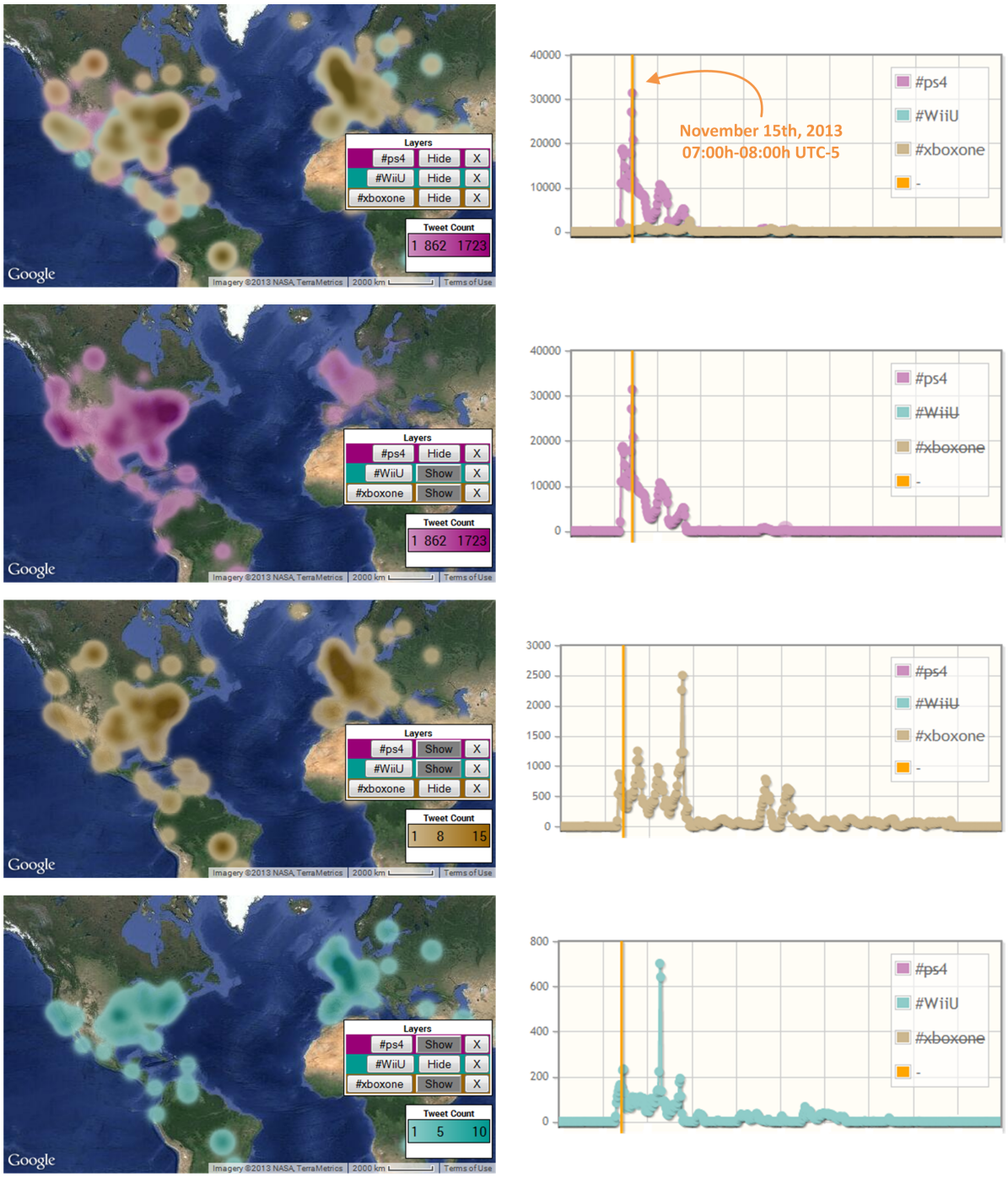


Figure 5: Heatmap-based as well as quantitative comparison of game console popularity.

tract from the collected tweets. Firstly, Figure 5 at a glance reveals the existence of large quantitative differences between the three tracked keywords. In the monitored time interval, the Wii U console garnered only a fraction of the attention that the Xbox One was able to accumulate, whose Twitter coverage in turn was outclassed by that of the PlayStation 4 by an order of magnitude. The fact that the experiment encapsulated the PlayStation 4's USA release date definitely contributed to this outcome. In particular, inspection of the captured tweet messages confirmed considerable hype build-up as the PlayStation 4 release approached. For the same reason, the PlayStation 4 tweets geo-spatially tended towards the USA.

Secondly, the volume diagrams show that Microsoft was able to pierce the PlayStation 4's Twitter hegemony exactly once in the course of the experiment. This achievement can be attributed to a clever marketing strategy: by retweeting a message from the official Twitter account of Xbox France, users could reveal the identity of the French Xbox One ambassador, an opportunity that was massively seized by fans. The resulting retweets primarily originated from Western Europe, and France in particular (not shown in Figure 5).

5 RELATED WORK

The principle of creating map mashups of the geographic sources of tweets has been considered by a number of commercialized web services. Examples include TweepsMap (http://tweepsmap.com/), Trendsmap (http://trendsmap.com/), Tweereal (http://tweereal.com/), Tweetping (http://tweetping.net/) and GlobalTweets (http://globaltweets.com/). The first maps (the home location of) the followers of a particular user's Twitter account, the second provides a real-time, localized mashup of currently trending Twitter themes, and the final three offer real-time geographic visualization of Twitter posts.

The academic literature also holds a number of articles that deal with deriving geo-spatial insights from Twitter data. Stefanidis et al. have proposed a framework to harvest and analyze ambient geographic information (i.e., not specified in terms of explicit coordinates) from tweets (Stefanidis et al., 2013). The iScience Maps tool targets behavioral researchers interested in exploiting Twitter for localized social media analysis purposes (Reips and Garaizar, 2011). The global concept of applying Twitter as a distributed sensor network to identify and locate events in the physical world has been successfully explored by a number of analogous research initiatives (Sakaki et al., 2010; Boettcher and Lee, 2012; Crooks et al., 2013; Takahashi et al., 2011); of particular relevance is the social pixel/images/video approach by Singh et al. that allows for Twitter-powered situation detection and spatio-temporal assessments (Singh et al., 2010). Field and O'Brien have investigated the application of cartographic principles to Twitter-powered map mashups (Field and O'Brien, 2010). Finally, the software architecture proposed by Oussalah et al. affords the deployment of geolocated services that are fueled by Twitter data (Oussalah et al., 2013).

All systems that have been cited in this section, both commercialized and academic ones, have their specific merits and feature sets. The TweetPos instrument exhibits functional overlaps with all of them. For example, the social pixel approach largely corresponds with our animated heatmap-based visualization solution. Some related tools even provide functionality that is missing in TweetPos. When for instance again looking at the social pixel framework, it incorporates an automated situation detection scheme and exploits domain semantics to autonomously recommend relevant control actions in response to detected events. However, the TweetPos tool exceeds every cited initiative in terms of the variety of analytical means it integrates and the synergistic benefits that stem from this holistic design. As an example, only a minority of the related systems grants insight in both historical and current tweet posting behavior. Also, the combination of a heatmap-based representation of the geographic intensity of topics, a tweet volume diagram, and dynamic means to inspect the textual contents of tweets fosters unprecedented deep mining of (the geo-spatial evolution of) Twitter contributions. A final example of a differentiating TweetPos feature is its layering mechanism and the opportunities in terms of comparative analysis it unlocks. Only the iScience Maps tool provides similar functionality, yet its comparison options are limited to exactly two configurations; in contrast, unlimited numbers of layers can be constructed in TweetPos.

6 CONCLUSIONS AND FUTURE WORK

SNSs have become prominent information channels in present-day society, as is manifested by the massive amounts of information that are shared and communicated through them. Given this quantitative overload, human operators benefit from tools that assist in transforming the constituting raw data into practical knowledge. This article has proposed *TweetPos*, a web service that provides exactly such assistive functions for the Twitter network, hereby allocating elevated attention to the geo-spatial characteristics of tweets. As the human mind is very adept at visual pattern recognition and at interpreting graphical data formats, TweetPos maximally invests in visual output modalities. The tool integrates and blends multiple complementary functions in order to yield a holistic solution for Twitter data analysis. Experimental results collected from two isolated test cases confirm this claim and prove the feasibility, effectiveness and added value of our work. In particular, it has been established that the TweetPos service succeeds in streamlining the ingestion, filtering, process-

ing, analysis and mining of tweeted information, and as such represents a valuable, highly versatile tool with cross-disciplinary application options.

Decision making logic, provisions for automated conclusion drawing and autonomous recommendation systems have deliberately been omitted from the current instantiation of the proposed tool, as we believe these tasks are more suited to human operators than to machines. As part of future research, we nonetheless plan to investigate whether the incorporation of computer-mediated aids might assist users in executing these actions more efficiently and swiftly. Potential supportive technologies include visual pattern recognition and edge detection algorithms to facilitate heatmap analysis, and linguistic processing frameworks to aid human operators in categorizing aggregated tweets on the basis of the textual message they convey. Another trajectory of future work is dynamic data delivery. In the current implementation, all tweet data pertaining to a particular query is transferred from the back-end server to the web browser in bulk. Although this design renders the TweetPos service highly responsive once all data has been downloaded, it also causes start-up delays to be high (i.e., they are directly proportional to the data set size). At the same time, network bandwidth utilization is suboptimal, as the client is likely to end up downloading data which the user will never inspect (or at least not in detail). We will therefore implement a demand-oriented transmission scheme in which relevant data is transmitted just-in-time (i.e., when it becomes needed). By doing so, we will be able to investigate the trade-off between service responsiveness and start-up delay, as well as the impact this balance has on the usage experience.

REFERENCES

140dev (2013). 140dev Streaming API Framework. Online, http://140dev.com/free-twitter-api-source-code-library/.

Boettcher, A. and Lee, D. (2012). EventRadar: A Real-Time Local Event Detection Scheme Using Twitter Stream. In *Proc. GreenCom 2012*, pages 358–367, Besançon, France.

Crooks, A., Croitoru, A., Stefanidis, A., and Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1):124–147.

Field, K. and O'Brien, J. (2010). Cartoblography: Experiments in Using and Organising the Spatial Context of Micro-blogging. *Transactions in GIS*, 14(s1):5–23.

Google Developers (2013a). Google Maps JavaScript API v3. Online, https://developers.google.com/maps/documentation/javascript/.

Google Developers (2013b). The Google Geocoding API. Online, https://developers.google.com/maps/documentation/geocoding/.

Oussalah, M., Bhat, F., Challis, K., and Schnier, T. (2013). A Software Architecture for Twitter Collection, Search and Geolocation Services. *Knowledge-Based Systems*, 37:105–120.

Pinto, N., Majaj, N. J., Barhomi, Y., Solomon, E. A., Cox, D. D., and DiCarlo, J. J. (2010). Human versus Machine: Comparing Visual Object Recognition Systems on a Level Playing Field. In *Proc. Cosyne 2010*, Salt Lake City, UT, USA.

Reips, U.-D. and Garaizar, P. (2011). Mining Twitter: A Source for Psychological Wisdom of the Crowds. *Behavior Research Methods*, 43(3):635–642.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. WWW 2010*, pages 851–860, Raleigh, NC, USA.

Singh, V. K., Gao, M., and Jain, R. (2010). Situation Detection and Control Using Spatio-temporal Analysis of Microblogs. In *Proc. WWW 2010*, pages 1181–1182, Raleigh, NC, USA.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting Ambient Geospatial Information from Social Media Feeds. *GeoJournal*, 78(2):319–338.

Takahashi, T., Abe, S., and Igata, N. (2011). Can Twitter Be an Alternative of Real-world Sensors? In *Proc. HCI International 2011*, pages 240–249, Orlando, FL, USA.

Twitter Developers (2013). Using the Twitter Search API. Online, https://dev.twitter.com/docs/using-search.

SHORT PAPERS

Experiences from and Attitudes towards Applying User Participation in Public e-Service Development

Jesper Holgersson and Eva Söderström
School of Informatics, University of Skövde, Skövde, Sweden
jesper.holgersson, eva.soderstrom@his.se

Keywords: e-government, Public e-Service Development, User Participation, Techniques for User Participation.

Abstract: It is evident that user participation is a vital component for successful public e-service development. However, it is also apparent that there is little guidance in e-government research about how user participation should be implemented in practice. Some high level guidelines can be found regarding user participation design schools but there is very little guidance in existing research regarding how these design schools can be implemented in practice. In this paper we have explored public administrations' experience of user participation, both in general systems development and in development of public e-services, in order to identify basic requirements that have to be fulfilled when implementing user participation in public e-service development. Thereafter we have applied these requirements on commonly used techniques to implement user participation in the light of three design schools: Participatory Design, User Centered Design, and User Innovation. Our results show that techniques to implement user participation in public e-service development must be adjusted to limited resources in terms of time and money as well as short development projects.

1 INTRODUCTION

Public electronic services (public e-services) , an instantiation of e-government, have become an increasingly adopted channel for the delivery of services used for interaction between public administrations and members of the society, such as citizens (Rowley, 2006). So far, the main arguments for implementing public e-services have been internal cost reductions achieved via decreased time and effort needed for handling previously manual tasks (Anthopoulos et al., 2007). This has led to a situation where public e-services have been designed from an internal perspective wherein external user considerations have been given only little attention (Axelsson et al., 2010). In order to find public e-services useful and beneficial, external users, such as citizens, must experience that they are served and satisfied when using the e-services. Otherwise they will go back to other more traditional channels when interacting with public administrations (Anthopoulos et al., 2007). Increased knowledge of citizen needs is essential for successful development of public e-services (Verdegem and Verleye, 2009) and by applying user participation in public e-service development the likelihood of positive effects on service usage is expected to increase, not only from the providing public administrations' point of view but also from a citizen perspective (Andersen and Henriksen, 2006).

However, user participation in public e-service development is challenging to put into practice (Axelsson et al., 2010) and many weaknesses in public e-service programs come from a failure to address needs from citizens. This is partly due to the voluntary nature of citizen participation in using public e-services (Saha, 2008) if compared to employed personnel, which can be obliged to participate (Albinsson and Forsgren, 2005).

There exist numerous methods and techniques within for how to implement user participation in systems development. In this work, we acknowledge a method as a recommended series of steps to follow, wherein each step may encompass one or several techniques used (Avison and Fitzgerald, 1995). Despite a vast number of available techniques, however, there is little or no knowledge about which techniques are the most efficient and effective in a certain public e-service development scenario (Maiden and Rugg, 1996) as well as how these techniques may be combined in order to obtain a comprehensive overview of citizen needs.

Furthermore, web applications such as public e-services often require a more rigorous requirements elicitation process due to the large number of potential end users (Escalone and Koch, 2004).

Government administrations need to implement user participation efficiently and accurately, but too little is currently known of how user participation is treated and adopted or not adopted by government administrations and what techniques are used when in the development process. The aim of this paper is to provide an analysis of current practice in the application of user participation techniques using three user participation design schools (User Centered Design, Participatory Design and User innovation) and an empirical study including public administrations in the form of government authorities, county councils and municipalities as a basis. Common techniques within the three schools will be analyzed and placed in an e-government context in terms of foundational requirements of user participation.

2 USER PARTICIPATION IN PUBLIC E-SERVICE DEVELOPMENT

User participation is a valuable and needed component in public e-service development. Jones et al. (Jones et al., 2007, p. 150) state that the "*key to the success of any e-government deployment is the citizen*" when they propose an agenda in order to better understand the approaches to citizen engagement. In fact, excluding users for example in the requirements engineering stage of public e-service development can be a problem (Folkerd & Spinelli, (2009). Early research on user participation focused largely on the user as an internal actor (Oostveen and van den Besselaar, 2004). More recently, the user concept has been widened to include external users, such as citizens and organizations that need to communicate with public administrations (Jansen, 2006).

User participation is a well-known concept in systems development (Mumford, 1981). However, only a limited number of studies discuss user participation in public e-service development, and even fewer focus on techniques for how user participation can be implemented. Karlsson et al. (2012) and Axelsson et al. (2010) both present challenges to applying user participation in public e-service development. Karlsson et al. (2012) analyze three user participation schools (User Centered Design – UCD, Participatory Design – PD, and User Innovation – UI) from a goal fulfillment perspective in order to assess to what extent they fit into a public e-service development perspective. Holgersson and Karlsson (Holgersson and Karlsson, 2011, Holgersson and Karlsson, 2012) build on these findings by empirically investigating to what extent the challenges exist in practice. Chutimaskul (2003) stresses the importance of user participation in requirements elicitation for public e-service development, but does not discuss how techniques for requirements elicitation may be applied. Van Velsen et al (2009) include such aspects in their citizen-centric approach towards end user RE for public e-services, including data collection techniques such as interviews and low-fidelity prototypes. Tsumaki and Tamai (2006) emphasize the importance of choosing the appropriate technique and the importance of carefully considering context and surrounding factors. However, neither of these studies gives advice for how to apply techniques for implementing user participation in the public e-service development context. This context differs from classical systems development where there is often a well-known and easily accessible user that can be obliged to participate in the development process. Some guidance exists for evaluating what school of user participation is the most preferable when applying user participation in public e-service development (Holgersson and Karlsson, 2011, Holgersson and Karlsson, 2012). There is only little guidance on how the user participation approaches should be implemented in practice in terms of more specific techniques, which hampers public administrations in mitigating these findings to public e-service development projects.

3 RESEARCH DESIGN

This study is based on two main data sources, a literature review on techniques to implement user participation and an interview study on user participation in public administrations. The literature review has been based on commonly used techniques in the three user participation design schools UCD, PD, and UI. For each of the design schools, searches regarding techniques to implement user participation have been made broadly, including journals in IS and HCI as well as relevant conference proceedings and books.

The empirical data collection of current practice in the application of user participation techniques

was based on semi structured interviews with public e-service providers representing municipalities, county councils and government authorities. In total, 22 interviews were performed, distributed over 6 municipalities (ranging in size from 5.000 to 134.000 citizens), 6 county councils and 10 government authorities. The respondents were selected based on their current involvement in public e-service development projects and had work titles such as project managers, CIOs, and business developers. The interviews were carried out either face to face or by telephone and took about 30 minutes. The primary goal of the interviews was to obtain qualitative data regarding public e-service providers' experiences of applying user participation in public e-service development. All interviews were recorded and transcribed. The material was then analyzed from the perspectives of what factors that affect or govern how user participation is implemented. The goal of the analysis was to identify relevant techniques to implement user participation, when they were applied or not and the motivations behind these choices. It should be noted that our focus is user participation in the development process of public e-services, not the end product in terms of the e-services themselves.

4 THREE DESIGN SCHOOLS FOR USER PARTICIPATION

There are numerous methods that support user participation. As an example, Muller et al. (1997) list 61 participatory methods, such as Joint Application Development (Wood and Silver, 1995) and ETHICS (Mumford, 1993). Aggregated on a higher level, three user participation design schools can be derived; User Centered Design (UCD), Participatory Design (PD), and User Innovation (UI) (Karlsson et al., 2012).

UCD is a design school introduced by Norman (Norman, 1982) in the late 1970's as a branch of the human computer interaction research field. Users and designers are here not seen as equal partners. The developers are the designers and the users are predominantly seen as passive advisors (Kling, 1977) who participates to a varied extent in the design process. Responsibilities and final decisions are still made by the developers (Gulliksen et al., 2003).

PD aims at improving how technology can assist users in performing their work (Kensing and Blomberg, 1998). The perhaps most radical approach to PD is the Scandinavian Design School. In this approach system developers and users are treated as equal partners (Karlsson et al., 2012) and both users as well as developers share responsibility to cooperate in a mutual dialogue (Carmel et al., 1993). This means that users must participate in decision making, either as advisors, as representatives or on a consensus basis (Mumford, 1981). It is the systems developers' responsibility to present a design that the users can respond to.

UI aims at capturing innovations from so called lead users (von Hippel, 1986). Lead users hold the key to innovative design proposals within the domain they are experts in (Kujala and Kauppinen, 2004). Put into practice, this means that the users identify problems as well as possible design solutions to the problems (von Hippel, 1986). The systems developers' main task is hence to capture lead users' ideas and designs and transform these into full-scale solutions in collaboration with the users.

Seen from a top down perspective, these three design schools represent different points on a floating scale wherein form and degree of participation varies, from passive forms of participation building on user representatives to more direct forms building on participation as advisors, consultants, and lead users (Bjerknes and Bratteteig, 1995, Cavaye, 1995). In this sense, UCD represents the least demanding form of participation whereas UI represents the most demanding form (Karlsson et al., 2012).

5 ANALYSIS

The analysis includes two parts: First, a literature review is made concerning commonly used techniques when implementing user participation in each of the three design schools (section 5.1); secondly, empirical findings are expressed with which the literature results will be compared (section 5.2). Our advices to public administrations are highlighted in italics.

5.1 Commonly Used Techniques in Literature

For UCD as well as for PD there are a number of techniques eligible for user participation. In many cases, UCD and PD are regarded as nearly equivalent concepts that often overlap with each other (Scandurra, 2007). This implies that it is generally difficult to identify techniques that are

unique for a particular design school. In many cases, the same technique is associated with both design schools, but at the same time, the same technique is applied differently. Thus, there is no 1:1 mapping between techniques and participation schools. As an example, focus groups are a technique that can be used for various purposes and in various forms. In UCD, focus groups are mainly used for aspects related to HCI (Gulliksen and Göransson, 2002), whereas they in PD are used for a larger spectrum of design activities (Bødker and Grønbæk, 1991).

The most commonly used techniques in UCD and PD are workshops in various formats (Bødker et al., 1991, Greenbaum and Kyng, 1991), observations (Folkerd and Spinelli, 2009), interviews (Gulliksen and Göransson, 2002, Oostveen and van den Besselaar, 2004, Bødker and Grønbæk, 1991), scenarios (Gulliksen and Göransson, 2002, Carroll et al., 2000), and various forms of prototyping (Bardram, 2000, Ehn and Kyng, 1991). UI is not associated with the same eligible techniques as for UCD and PD in literature. The main reason is that UI exists in several variants, such as end user development, where end users construct their own information systems (Taylor et al., 1998). Therefore, UI does not have structured techniques for its implementation as UCD and PD have. *Public administrations should be aware of UI not always being a first-hand choice due to its demands in terms of resources and time.* The main emphasis in UI lies on the user working together with the designers in an iterative fashion as an integrated part of the users ordinary work duties (Kujala and Kauppinen, 2004).

5.2 Empirical Findings: Public Administrations' View on Public e-Services and User Participation

The analysis of the interviews reveals differences in how public administrations view e-services and user participation in the development thereof. Despite these differences, there is a consensus on what features that affect user participation in public e-service development. During the analysis of the interviews, it became evident that most public authorities and county councils have established ways of working with user participation. As one respondent puts it: "*We use those user groups that are affected by the service that we will present and look at what they think and what they would like to change. We cannot sit here and think that we know best, because we do not. Those who will use it know best*". This is an interesting comment. The respondents seem aware of their own limitations. However, it is important to remember that *users sometimes need guidance in what they want as well as in what is possible to achieve and what is not*. When it comes to municipalities, user participation varies. Smaller municipalities (less than 10.000 residents) seem not to apply user participation at all, mostly because they do not see any basis for implementing public e-services and no economic gains to be made. The following quotes exemplify the situation: "*We see no gains to start using e-services since 1) there is no demands from the citizens, and 2) there is no saving but only cost increases with e-services*", "*If we take for example the application for alcohol licensing it is perhaps four such cases per year to handle. Developing an e-service for this will never pay off*". Two immediate questions arise from the first quote: How do they know there are no demands from the citizens, and how did they calculate the cost increases versus lack of savings? At best, the organization has made a thorough investigation of both issues, but it is also possible that they base their decision to develop or not develop an e-service on assumptions. If the latter is the case, *there is a clear need for practical guidelines to aid in and guide through the development decisions*. Larger municipalities (more than 50.000 residents) apply user participation most sparsely despite the fact they have many public e-services. These larger municipalities all wish to devote more attention to their external end users: "*User participation is something that we do not work enough with. It feels a bit awkward to ask the citizens what they think. We have been a bit cowardly there and instead went to our administrations that get a lot of impressions and feedback from citizens*", "*We have not had the opportunity to have the users in development but I think it would be great to test this. As it is today, it is the administrative officers' needs and what they believe the citizens need that determines*". Both quotes illustrate how the users are part of development indirectly and that there is a will to try to include them, but also that there is a fear for doing so. This fear may stem from a lack of knowledge of how to address and include citizens. This suggests *a need for practical guidelines for how to apply the techniques*.

5.3 Features that Affect User Participation

When discussing important features regarding techniques to implement user participation public

administrations in general experience the same problems, regardless of how experienced they are in putting user participation into practice. Techniques for user participation in public e-service development must take into consideration that the time for development in many cases in sparse, or as two respondents put it: "*We always are in a hurry. It takes time to make the applications since there are so many other related systems. This requires extensive testing in order to make sure that everything works correctly. We start to work on the following year's version of the application in May and then we have a year until the next version must be ready to use*", "*Since our project is a EU-project financed for three years we have to stay within these time frames. That means that we have to run pretty fast*". Slimmed resources in terms of time and economy clearly are delimiting factors for user participation in public e-service development, as the following quotes illustrate: "*For obvious reasons, time and money limits how we can work with the citizens*", "*It is actually a question of resources to cope with doing it too besides everything else*", "*We probably would have worked in a different way if we had had more time*". Time and money are influencing factors. However, there may be an underlying assumption that user participation requires time without an actual testing of or investigation into whether or not this is true. Again, we see the *need for practical guidance to developers in order to make more informed decisions*.

5.4 Summarizing Discussion

In summary, in order to apply user participation in public e-service development one must take into account that development cycles often are short and time to spend with external end users is limited. Based on our results, we give the following advice to public administrations:

- *UI is not always a first-hand choice due to its demands in terms of resources and time.*
- *Users sometimes need guidance in what they want, what is possible to achieve, and what is not.*
- *There is a clear need for practical guidance for making informed development decisions, and how to apply the techniques.*

Public administrations often experience limited resources in terms of time and money available for external user engagements. However, these experiences may be based on assumptions without thorough investigations of their truth value. More practical guidance is needed to help investigating these situations.

6 CONCLUSIONS

User participation in public e-service development is a matter that most public administrations wish to devote more resources to. However, scarce resources, little time available and short development projects are limiting factors that must be taken into consideration when discussing the issue. The limitations mean that not all techniques available for putting user participation into practice are possible to implement. None of the techniques discussed in chapter 5.1 are disqualified from usage in public e-service development, however. What matters is *how* a certain technique is implemented. For example, workshops are an umbrella term for a number of techniques that may be used for user participation (Nuseibeh and Easterbrook, 2000). A workshop can be executed in variety of ways, such as focus groups (Kuniavsky, 2003), Joint Application Development (JAD) (Carmel et al., 1993), Interactive prototyping (Bødker and Grønbæk, 1991) etc. Furthermore, each of these ways can be executed in different manners, such as in close face to face iterative group sessions or as individual sessions held via Skype and similar (Sanders et al., 2010) depending on what design school that is considered. Another example is the usage of prototypes, which will look a lot different depending on the design school chosen. In UCD, prototypes will be used mostly as a technique to evaluate usability and functionality from a user perspective. In PD, prototypes will be used as a means for designers and users to actively elaborate during development. Design decisions are joint decisions by users and designers. In UI, prototypes will be used as a means for lead users to demonstrate potential solutions to identified problems for system developers This large variety in technique execution makes it difficult to recommend a specific technique for a specific matter. All techniques discussed in chapter 5.1 fail to fulfill the basic requirements posed by public administrations. At the same time, if executed in another fashion, every technique is eligible. In conclusion, the *how* is more important than the *what*. Our findings are in line with previous research regarding citizens and business employees willingness and ablitity to particpate in public e-service development (Holgersson and Karlsson, 2011, Holgersson and Karlsson, 2012) where it is concluded that it is important for public

administrations to use techniques that are efficient and less time consuming in order to get business employees and citizens interested in participating.

Even so, there is a knowledge gap concerning how existing techniques for implementing user participation work and how they may be implemented when developing public e-services. This paper has originated from three user participation design schools and the analysis indicates that all participation techniques are eligible but must be adjusted with respect to time efficiency and to the context that most public administrations belong to. They must also follow the design principles for each of the design schools and this means that each technique will be applied differently. Future work is hence needed to develop concrete and practical guidelines that will help public administrations to develop useful and usable e-services in collaboration with the intended users.

ACKNOWLEDGEMENTS

This work has been conducted within the DOME research project, sponsored by the Swedish Agency for Innovation Systems (Vinnova).

REFERENCES

Albinsson, L. & Forsgren, K. 2005. Co-Design Metaphors And Scenarios - Two Elements In A Design Language For Co-Design. *Lap.* Kiruna.

Andersen, K. V. & Henriksen, H. Z. 2006. E-Government Maturity Models: Extension Of The Layne And Lee Model. *Government Information Quarterly,* 23**,** 236-248.

Anthopoulos, L.G., Siozos, P. S. & Tsoukalas, I. A. 2007. Applying Participatory Design And Collaboration In Digital Public Services For Discivering And Re-Designing E-Government Services. *Government Information Quarterly,* 24**,** 353-376.

Avison, D. & Fitzgerald, B. 1995. *Information Systems Development: Methodologies, Techniques And Tools,* Uk, Mcgraw-Hill.

Axelsson, K., Melin, U. & Lindgren, I. 2010. Exploring The Importance Of Citizen Participation And Involvement In E-Government Projects - Practice, Incentives And Organization. *Transforming Government: People, Process And Policy,* 4**,** 299-321.

Bardram, J. E. 2000. Scenario-Based Design Of Cooperative Systems: Re-Designing An Hospital Information System In Denmark. *Group Decision And Negotiation,* 9**,** 237-250.

Bjerknes, G. & Bratteteig, T. 1995. User Participation And Democracy. A Discussion Of Scandinavian Research On Systems Development. *Scandinavian Journal Of Information Systems,* 7**,** 73-98.

Bødker, S., Greenbaum, J. & Kyng, M. 1991. Setting The Stages For Design As Action. *In:* Greenbaum, J. & Kyng, M. (Eds.) *Design At Work.* New Jersey: Lawrence Erlbaum.

Bødker, S. & Grønbæk, K. 1991. Design In Action: From Prototyping By Demonstration To Cooperative Prototyping. *In:* Greenbaum, J. & Kyng, M. (Eds.) *Design At Work.* New Jersey: Lawrence Erlbaum.

Carmel, E., Whitaker, R. D. & George, J. F. 1993. Pd And Joint Application Design: A Transatlantic Comparison. *Commun. Acm,* 36**,** 40-48.

Carroll, J. M., Chin, G., Rosson, M. B. & Neale, D. C. 2000. The Development Of Cooperation: Five Years Of Participatory Design In The Virtual School. *Proceedings Of The 3rd Conference On Designing Interactive Systems: Processes, Practices, Methods, And Techniques.* New York City, New York, Usa: Acm.

Cavaye, A., L, M. 1995. User Participation Is Systems Development Revisited. *Information And Management,* 28**,** 311-323.

Chutimaskul, W. 2003. E-Government Requirement Elicitation. *In:* Tranmuller, R. (Ed.) *Egov.* Prague, Czech Republic: Springer.

Ehn, P. & Kyng, M. 1991. Cardboard Computers: Mocking-It-Up Or Hands-On The Future. *In:* Greenbaum, J. & Kyng, M. (Eds.) *Design At Work: Cooperative Design Of Computer System.* New Jersey: Lawrence Erlbaum Associates.

Escalone, M. & Koch, N. 2004. Requirements Engineering For Web Applications - A Comparative Study. *Journal Of Web Engineering,* 2**,** 193-212.

Folkerd, C. & Spinelli, G. 2009. User Exclusion And Fragmented Requirements Capture In Publicly-Funded Is Projects. *Transforming Government: People, Process And Policy,* 3**,** 32-49.

Greenbaum, J. & Kyng, M. 1991. Introduction: Situated Design. *In:* Greenbaum, J. & Hillsdale, M. (Eds.) *Design At Work: Cooperative Design Of Computer Systems.* New Jersey: Lawrence Erlbaum Associates.

Gulliksen, J. & Göransson, B. 2002. *Användarcentrerad Systemdesign,* Lund, Studentlitteratur.

Gulliksen, J., Göransson, B., Boivie, I., Blomkvist, S., Persson, J. & Cajander, Å. 2003. Key Principles For User-Centred Systems Design. *Behaviour And Information Technology,* 22**,** 397-409.

Holgersson, J. & Karlsson, F. 2011. Exploring Citizens' Possibilities To Participate In Public E-Service Development. *In:* Chappelet, J.-L., Glassey, O., Janssen, M., Macintosh, A., Scholl, H. J., Tambouris, E. & Wimmer, M. A. (Eds.) *Egov 2011* Delft: Trauner Verlag.

Holgersson, J. & Karlsson, F. 2012. Understanding Business Employees' Conditions For Participating In Public E-Service Development. *In:* Pries-Heje, J., Chiasson, M., Wareham, J., Busquets, X., Valor, J. & Seiber, S. (Eds.) *Ecis 2012.* Barcelona.

Jansen, A. 2006. Strategies For User Involvement In Egovernment Projects: What Can Be Learned From The Scandinavian Is Tradition? *In:* Følstad, A., Artman, H. & Krogstie, J. (Eds.) *Nordichi'06 Workshop On User Involvement And Representation In E-Government Projects.* Oslo.

Jones, S., Hackney, R. & Irani, Z. 2007. Towards E-Government Transformation: Conceptualising "Citizen Engagement". *Transforming Government: People, Process And Policy,* 1**,** 145-152.

Karlsson, F., Holgersson, J., Söderström, E. & Hedström, K. 2012. Exploring User Participation Approaches In Public E-Service Development. *Government Information Quarterly,* 29**,** 158-168.

Kensing, F. & Blomberg, J. 1998. Participatory Design: Issues And Concerns. *Computer Supported Cooperative Work,* 7**,** 167-185.

Kling, R. 1977. The Organizational Context Of User-Centered Software Designs. *Mis Quarterly,* 1**,** 41.

Kujala, S. & Kauppinen, M. Identifying And Selecting Users For User-Centered Design. *In:* Raisamo, R., Ed. The Third Nordic Conference On Human-Computer Interaction, 2004 Tampere, Finland. Association For Computing Machinery, 297 - 303.

Kuniavsky, M. 2003. *Observing The User Experience: A Practitioner's Guide To User Research,* San Fransisco, Morgan Kaufmann.

Maiden, N. & Rugg, G. 1996. Acre: Selecting Methods For Requirements Aquisition. *Software Engineering Journal,* 11**,** 183-192.

Muller, M. J., Hallewell Haslwanter, J. & Dayton, T. 1997. Participatory Practices In The Software Lifecycle. *In:* Helander, M., Landhauer, T. K. & Prabhu, P. (Eds.) *Handbook Of Human-Computer Interaction.* Amsterdam: Elsevier.

Mumford, E. 1981. Participative Systems Design: Structure And Method. *Systems, Objectives, Solutions,* 1**,** 5-19.

Mumford, E. 1993. The Ethics Approach. *Communications Of The Acm,* 36**,** 82.

Norman, D. Steps Toward A Cognitive Engineering: Design Rules Based On Analyses Of Human Error. *In:* Nichols, J. A. & Schneider, M. L., Eds. Human Factors In Computing Systems, 1982 New York. Association For Computing Machinery, 378-382.

Nuseibeh, B. & Easterbrook, S. 2000. Requirements Engineering: A Raodmap. *Proceedings Of The Conference On Hte Future Of Software Engineering (Icse).*

Oostveen, A.-M. & Van Den Besselaar, P. 2004. From Small Scale To Large Scale User Participation: A Case Study Of Participatory Design In E-Government Systems. *The Participatory Design Conference.* Toronto, Canada.

Rowley, J. 2006. An Analysis Of The E-Service Literature: Towards A Research Agenda. *Internet Research: Electronic Networking Applications And Policy,* 16**,** 879-897.

Saha, P. 2008. *Government E-Service Delivery: Identification Of Success Factors From Citizens' Perspective.* Phd, Luleå University Of Technology.

Sanders, E. B.-N., Brandt, E. & Binder, T. 2010. A Framwork For Organizing The Tools And Techniques Of Participatory Design. *The 11th Biennial Participatory Design Conference (Pdc).* Sydney, Australia: Acm.

Scandurra, I. 2007. *Building Usability Into Health Informatics - Development And Evaluation Of Information Systems For Shared Homecare.* Phd, Uppsala University.

Taylor, M. J., Moynihan, E. P. & Wood-Harper, A. T. 1998. End-User Computing And Information Systems Methodologies. *Information Systems Journal,* 8**,** 85-96.

Tsumaki, T. & Tamai, T. 2006. Framework For Matching Requirements Elicitation Techniques To Project Characteristics. *Software Process Improvement And Practice*.

Verdegem, P. & Verleye, G. 2009. User-Centered E-Government In Practice: A Comprehensive Model For Measuring User Satisfaction. *Government Information Quarterly,* 26**,** 487-497.

Von Hippel, E. 1986. Lead Users: A Source Of Novel Product Concepts. *Management Science,* 32**,** 791-805.

Wood, J. & Silver, D. 1995. *Joint Application Development,* New York, John Wiley & Sons Inc.

The Co-retweeted Network and Its Applications for Measuring the Perceived Political Polarization

Samantha Finn, Eni Mustafaraj and Panagiotis T. Metaxas
Department of Computer Science, Wellesley College, Wellesley, MA02481, U.S.A.
{*sfinn, emustafa, pmetaxas*}*@wellesley.edu*

Keywords: Social Media Analytics, Social Networks, Twitter, Big Data, Social and Legal Issues, Human Computation.

Abstract: This paper introduces a novel network, the co-retweeted network, that is constructed as the undirected weighted graph that connects highly visible accounts who have been retweeted by members of the audience during some real-time event. Like bibliographics co-citation used to indicate that two papers treat a related subject matter, co-retweeting is used to indicate that two accounts present similar opinions in an online discussion. Thus, the co-retweeted network can be seen as a form of consulting the opinion of the crowd that is following the discussion about the similarity (or difference) of positions expressed by the highly visible accounts. When applied on political conversations related to some event, the co-retweeted network enables the measurement of the polarity of political orientation of major players (including news organizations) based on the views of the audience. It can also measure the degree of polarization of the event itself.

1 INTRODUCTION

Presidential debates in the United States are very important political events. Their television audience consistently ranks among the largest of the year (second only to the famous "Super Bowl", the championship game of US Football). However, these spectators are no longer passive: they increasingly use the web as a platform for further engagement. As studies from Pew Research have shown, 1 in 10 spectators in presidential debates is a "dual-screener" [1]. Very often the second screen is Twitter, where running commentary of live-televised events is at its liveliest.

But studying such lively online discussions has a significant obstacle: data size. During the 2012 US presidential race President Barack Obama and challenger Mitt Romney held three debates on October 3rd, 16th and 22nd. These debates generated respectively 10.3 million[2], 7.2 million[3], and 6.5 million[4] tweets in a time span of approximately 90 min. each.

Currently, only Twitter (the company) is able to make sense of such big data by creating simple histogram-like data aggregations that spike during the debates (charts can be found in footnotes 2, 3, 4). By establishing a correspondence between what was being said during the debate and the number of tweets per minute mentioning those words (and other debate-related hashtags), Twitter can quantify how moments during the live event affected the tweeting public.

While this is a good way to summarize the major issues tweeted during an event, one might be interested not only in what is being tweeted, but also who is tweeting it and why. In addition, there are other questions that often occupy the public discourse during political discussions: *How polarized were the political postings online? Did the major news organizations take sides in favor of one or the other candidate? And did the supporters of one discussant were more vocal than the supporters of the other?*

News organizations answer these questions by using human experts who try to make sense of the data, preferably while the event is developing. However, no matter how well trained, human experts will hardly be able to keep track of the avalanche of data coming at a speed of 100,000 tweets/minute. Our approach instead is to crowd source the solution by using "human computation" (von Ahn et al., 2006) in the form of analyzing the independent decisions made by large groups of users in parallel. In Twitter, this takes into account retweets[5] and favorites.

[1] Social media a hotbed of political debate, engagement – for the good? http://to.pbs.org/1bCvUoJ

[2] Dispatch from the Denver debate, http://bit.ly/1k15reF

[3] Twitter at the Town Hall Debate, http://bit.ly/1kWkBj6

[4] The Final Presidential Debate, http://bit.ly/19Av3VM

[5] In this paper, when we refer to "retweets" we mean verbatim retweets created by the clicking of the retweet button, not the manually created modified tweets.

In an effort to take advantage of this human computation, we propose the construction of a novel network, the co-retweeted network of Twitter accounts. Our insight is that the co-retweeted network displays some dynamics of Twitter conversations during a certain event as revealed by the repeated and independent behavior of the millions of users following the event. We define as *co-retweeting* the act of a single user retweeting two –or more– different accounts. These acts are used to create undirected weighted edges between the co-retweeted accounts in the network. The more users retweet these two accounts, the more weight the edge gains. And there can be a lot of retweets to include in the network construction. In fact, the data we collected during the presidential debates show that more than 50% of the tweets sent are retweets. However, some tweets will have greater visibility, receiving in the order of several thousand retweets, while the vast majority of them will only receive one or two retweets. We use this information to create a co-retweeted network that captures how the wider public views the major participants in the discussion. An example visualizing the second presidential debate using the co-retweeted network is shown in Figure 1. To the best of our knowledge, this has not been previously studied in the Twitter-related research.

We claim that the co-retweeted network is able to measure the perceived political orientation of the major players by recording the retweets of the audience on Twitter. Co-retweeting can be compared to the concept of *co-citation* (Small, 1973). The latter is a well known bibliometrics measurement used by librarians to determine whether two published works may be treating similar subjects or not. When there are a lot of works co-citing two papers, A and B, this is considered evidence that A and B treat related subjects. In this respect, it reflects the opinion of many authors and thus represents a better indicator of subject similarity. The more co-citations two documents receive, the more likely they are semantically related. (In fact, reviewers of papers enforce that by demanding the inclusion of missing co-citations.)

We use co-citation in the same spirit but, in this paper, in a more restricted domain: During a political event, when lots of Twitter users retweet two accounts, U_A and U_B, we consider it as evidence that U_A and U_B are taking politically similar positions. The more co-retweeting the tweets of U_A and U_B receive, the more likely they are politically related. This fact can lead to some concrete measurement on how polarized two accounts are and, when considering the overall political interaction, how polarized the event is. Effectively, we are able to compute the answers to the three questions we posed above.

The remaining of this paper is organized as follows. After describing prior work related to the issues treated in this paper, we define the co-retweeted network and describe its construction. Then we present the applications we consider in this paper, measuring the degree of political polarity during an online event, and measuring the polarity of the major players participating in the event. We end with our conclusions.

1.1 Prior Work

Co-citation was first introduced by Henry Small (Small, 1973) and is considered as a basic concept in Bibliometrics. The histogram-like data aggregations mentioned in the introduction was first applied by researchers during a debate for the 2008 US Elections (Diakopoulos and Shamma, 2010).

We introduced the concept of co-retweeting in (Finn and Mustafaraj, 2013), with a possible application for recommending relevant tweets during an event. The domain of recommendation and specifically the algorithms for item-based collaborative filtering are at the foundation of our method for building the co-retweeted network, which follows the principles of building the item-item similarity matrix with co-rated items (Sarwar et al., 2001). In the current paper, we expand the definition of the co-retweeted network and introduce new applications for it.

The first application, measuring opinion polarization for an issue or topic, is a research question that is usually treated in the political sciences literature. A review of the literature on this question (Fiorina and Abrams, 2008) concluded that the "American public as a whole is no more polarized today that it was a generation ago". While these studies focus on the public as a whole, we are posing this question in the context of the social media participants, which might not be representative of the entire public but still can command certain influence over the public discourse and its coverage in the traditional media. Furthermore, since Twitter has a global userbase, we can measure polarity in political discourse in other countries (e.g. Germany) as well.

The second application of the co-retweeted network is about calculating an account's polarity and using it to measure the media bias (as perceived by the Twitter audience). Researchers have always been concerned with media bias for a long time (for an overview refer to (Prior, 2013)) and different research fields use various methodologies to measure it. A somewhat related approach that makes use of Twitter, but doesn't calculate media bias (it calculates the political preferences of a (media) organization's Twit-

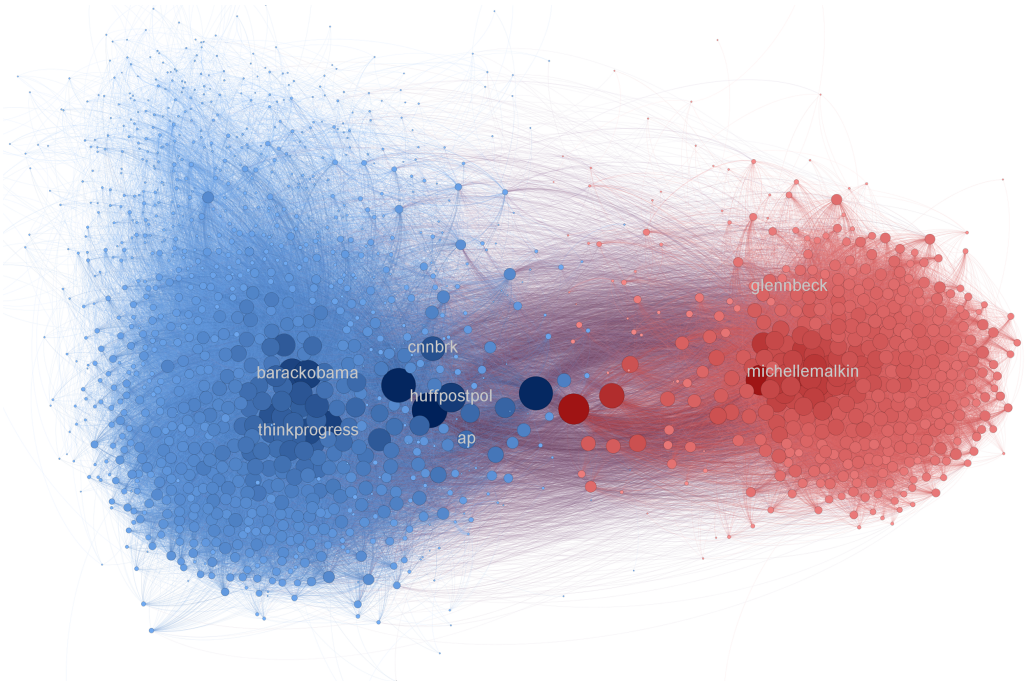


Figure 1: The network visualization for the 1,500 most co-retweeted accounts during the second debate. It turns out that, in this data set, the nodes colored blue by the community detection algorithm belong to accounts with a liberal political orientation and the nodes colored red to accounts with conservative political orientation. The labels indicate the location of some of the more visible accounts belonging to news organizations and political entities.

ter followers instead) can be found in (Golbeck and Hansen, 2014). Their method relies on the static relation of "followship". Our method instead is dynamic in the sense that every co-retweeted network is event-specific and may display different bias. Such specificity cannot be captured through static relations.

Additionally, we use the account polarity method to infer the polarity of all (engaged in retweeting) accounts in the audience to answer the question of whose supporters were more active and vocal during an event. This problem has certain similarities to the task of classifying Twitter users based on their political orientation as described in (Conover et al., 2011) and (Cohen and Ruths, 2013), however, our approach differs significantly from theirs. The cited research uses machine learning to classify users who have been represented by a set of features (mostly hashtags and other metadata from their tweets). We don't use the text of the tweets in our approach. Instead, the polarity (or orientation) is calculated based on the characteristics of the co-retweeted network, as well as the retweeting behavior of the users. A limitation of our method is that it cannot be used for users who didn't retweet. However, previous research has already established that the most vocal supporters engage heavily in retweeting (Mustafaraj et al., 2011; Jungherr et al., 2012).

2 THE CO-RETWEETED NETWORK

2.1 Data Collection

Collecting data from Twitter is a relatively easy task, due to the different APIs offered by Twitter. However, being sure that one has collected all the relevant or existing data is difficult to evaluate. The only way to be sure is to have access to the Twitter Firehose API, which is usually unreasonably expensive for academic researchers. Currently, researchers use the public APIs (Streaming, Search, REST) by adopting one (or a combination) of the following three methods (Gerlitz and Rieder, 2013):
1) a selection of topic-relevant hashtags and keywords which is fed to the Streaming or Search API,
2) a network of users and their followers (snowball sampling) for which the REST API collects all tweets,
3) metadata features (location, language, etc.) which are also fed to the Streaming or Search API.

For our study, we use the first method: a selection of topic-relevant hashtags and keywords. Two are the major problems with this method: a) the set of preselected keywords and hashtags might not capture the entirety of the discussions about a topic, and b) once

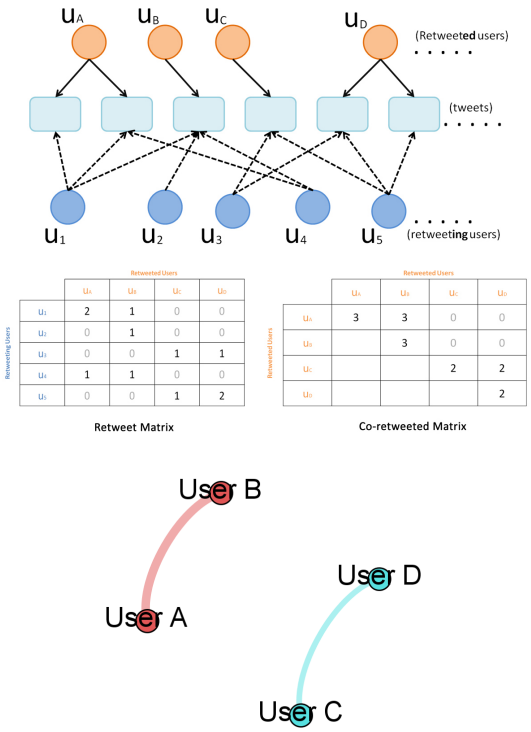


Figure 2: (Top) The process of generating the co-retweeted matrix. The graphic at the top displays the activity of users $U_A, U_B, \ldots$ whose tweets (blue rectangles) are being retweeted by other users $U_1, U_2, \ldots$. On the middle left is the corresponding retweet matrix and on the middle right the symmetric co-retweeted matrix. The main diagonal of the co-retweeted matrix shows the number of users retweeting a certain account, and the upper diagonal cells contain the number of times two accounts were co-retweeted. (Bottom) The co-retweeted network produced by the matrix above.

a topic of conversation becomes popular, Twitter applies rate limitations. All data collections have to deal with the first problem, while the second one applies only to conversations that become dominant (such as the US presidential debates discussed in this paper).

Our strategy for dealing with the first problem was the following: during the first presidential debate (Oct. 3, 2012) we used the keywords "obama" and "romney" to collect tweets mentioning the two contenders. The text of tweets was then processed to find the set of most common keywords and hashtags, which then was used for the two successive debates. Additionally, before every new debate we updated the list with hashtags being promoted before the TV broadcast such as #cnndebate, #lynndebate, etc. This strategy still doesn't ensure that we will be able to capture the whole conversation, but, for the purposes of our method, we don't need all tweets, we need the tweets that receive the most retweets. Previous research (Suh et al., 2010) has established that "a tweet with hashtags is more likely to get retweeted". By using the most popular hashtags around which the large community has coalesced, we are making the assumption that users who want to be retweeted will be using these agreed upon hashtags. Furthermore, our previous experience with collecting election data from Twitter (Mustafaraj et al., 2011) has shown that Twitter users who play the curator role, will add the appropriate hashtags before retweeting a message relevant to a conversation.

The problem of Twitter rate limits appears whenever the conversation about a topic or event exceeds the 1% of the whole Twitter volume. The presidential debates, which often averaged at 100,000 tweets/minute[6] were some of the most-tweeted events in the Twitter history (the first debate was the most-tweeted event ever). Thus, we know that we were able to capture only a small fraction of the sent tweets (in fact, Twitter includes a message in the Streaming API to indicate how many tweets were missed since the latest received tweet, e.g., `"limit": "track": 1234` , and we can calculate how many tweets we missed). Since we are mostly interested in popular retweets, one way to mitigate the rate limitation during the live TV coverage of the debates was to continue the data collection several hours beyond the event completion. This makes sure that we get retweets of content that users who didn't follow the live debate found interesting afterwards. We then tested that the 1,500 tweets with most retweets in our sample are a good representation of most retweeted content by the following procedure: in a later date, we recollected the top 3,000 tweets from our dataset, checked their retweet count, and ranked them based on this number. By comparing these two rankings, we found that our sample has 83% of the top 500 most retweeted tweets, 74% of top 1000, and 71% of top 1500, offering a good coverage, despite the limited sampling. Finally, recent research on comparing the Twitter Streaming API to the Firehose (Morstatter et al., 2013) has indicated that for network-level measures (such as centrality, etc.) there is correlation between the results of the two datasets. So, while our data collection method might not be perfect, we have taken into account all available measures to overcome the known problems.

2.2 Co-retweeted Network Construction

From the tweets collected as described above, we focus only on tweets with retweet information using the `retweeted_status` field included in the JSON rep-

[6]Tracking the #Debates: From Big Bird to Bayonets http://bit.ly/1dhigZ9

resentation of a tweet. Such information allows us to create pairs of the original message sender and the retweeter. This information is necessary for creating the co-retweeted network and this process is summarized in Figure 2. We create a matrix of *retweet* relationships, where each row represents a user, and the columns represent users who have been retweeted. From this matrix we obtain the *co-retweeted matrix*, containing only the users who were retweeted (the columns in the retweet matrix). Note that the co-retweeted matrix is symmetric, corresponding to a weighted undirected network, and we show only its upper diagonal.

Each co-retweeted matrix entry represents how many times the two users have both been retweeted by other users. For example, referring to Figure 2, in the retweet matrix, Users U_A and U_B have both been retweeted by a third user, User U_1. The entry for U_A and U_B in the co-retweeted matrix is incremented by 1. The same is true for U_4 who has also retweeted both U_A and U_B). Using this matrix we create the co-retweeted network of the retweeted accounts, where each user in the rows and columns of the matrix becomes a node, with an edge between them weighted by the number of users who have retweeted both of them. Accounts U_i, for numeric i, who have not been retweeted are ignored in the co-retweeted matrix.

2.3 Visualization of the Oct 16 Debate

Figure 1 shows the visualization[7] of the top 1,500 most co-retweeted accounts built using about 1.3 million tweets that we collected as described. This visualization was created using Gephi[8] that was chosen for convenience as it includes several useful implementations of community detection, force-directed layout and influence calculations. Other implementations are possible, of course, but using a comprehensive set of implementations was beyond the scope of this paper. After all, using comparable implementations the essence of the co-retweeted network should not change. The co-retweeted network in Figure 1 is layed out using Gephi's ForceAtlas2 implementation (Gephi, 2010) where nodes are attracted to each other based on the cardinality of their connections (Fruchterman and Reingold, 1991).

We notice two very distinct groups, with only a few nodes bridging the gap between them. The colors used in the visualization make these two groups more apparent and are based on the Louvain community finding algorithm (Blondel et al., 2008) which computes groups of nodes containing more edges within the group than outside the group. The size and darkness of each node is based on influence in the network indicated by their eigenvalue centrality value. By comparison to Figure 1, Figure 3 shows the retweet network of the top 1,500 accounts which is highly disconnected. The giant component contains only 550 connected nodes; the remaining nodes are not connected to it and are shown on the periphery. This is not surprising, since the most retweeted nodes do not retweet others, making the retweet network of the top accounts not very useful. Details about the datasets are listed in Table 1.

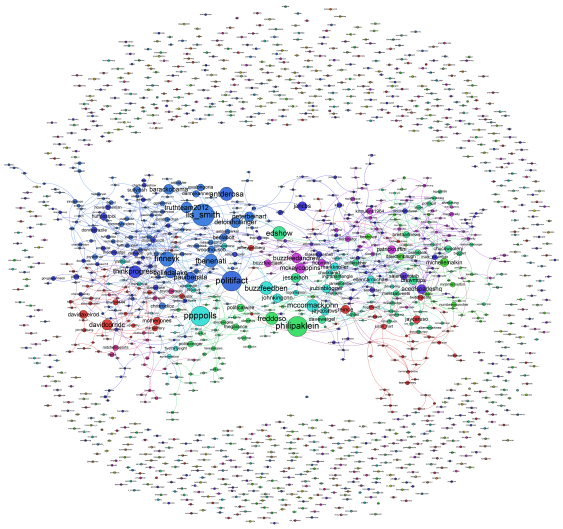


Figure 3: The retweet network of the 1,500 most retweeted users during the debates is not very useful in making sense of the online activity. The central component contains the 550 connected nodes and 6 communities; the remaining components are shown on the periphery. (The full retweet network –not shown here– is also not as useful.).

3 MEASURING THE PERCEIVED POLITICAL POLARIZATION

Next we describe a couple of applications for the co-retweeted network. First we describe the insight gained by observing the co-retweeted networks of the two presidential debates. Then we define the measurement "degree of the network polarity" and a method to compute the individual polarity of any member participating in a polarized network.

The visualizations for both presidential debates display two highly separated communities that break

[7]For an interactive version of the Oct. 16 debate, visit http://bit.ly/1d3MRJM and for the Oct 22 debate visit http://bit.ly/1fWajvH. Due to the size of the network and depending on the available bandwidth, loading the page may take a minute.

[8]The Gephi Visualization Tool, http://gephi.org

Table 1: Summary of the two debate datasets, which were used to create the co-retweeted networks. Users contributing refers to the number of users who retweeted at least two of the top 1,500 users (in this way they contribute at least one edge to the network). Retweets contributing refers to how many times the top 1,500 users were retweeted by the contributing users.

	Users	Tweets	Retweets	Nodes	Edges	Modularity	Contrib. Users	Contrib. RTs
Oct. 16	615,376	820,746	423,802	1,498	92,196	0.440	24,719	79,985
Oct. 22	646,559	1,074,588	582,526	1,496	126,252	0.434	45,493	179,978

down over political lines, with popular news media accounts bridging the gap between them. The distance between the two groups, determined by the force-directed algorithm, is a result of their polarization. The accounts in the blue group on the left of the network are largely liberal leaning politicians and popular bloggers (e.g. @barackobama, @thinkprogress), while to the right the red consists of conservative accounts (e.g. @glennbeck, @michellemalkin). Between these two groups are media accounts, which are divided between red and blue. Some of these accounts are claiming to be politically unbiased. However, many mainstream news accounts (@cnnbrk, @huffpostpol, @ap) in our event network are classified closer to the blue community.

Previous research has demonstrated how political social media is polarized in the two political orientations (Adamic and Glance, 2005). However, here the situation is different. The links are not created as a result of the actors actively connecting to each-other. In fact, although these are popular accounts, some of whom received thousands of retweets during the debates, they are not active in retweeting each other. In fact, the corresponding retweet graph (Fig. 3) is disconnected. Its giant component shows 6 communities which are not obviously distinguishable, and only the largest two show obvious political orientation. The co-retweeted network bypasses the sparsity of the retweet network by utilizing the hundreds of thousands of retweets made by normal users in order to form connections between the most popular accounts, and by doing this reveals the perceived orientation of the popular accounts. We make use of the assumption (Metaxas and Mustafaraj, 2010; Conover et al., 2011; Wellesley Trails Group, 2014) that the majority of users are not likely to retweet something that disagrees with their political views. Therefore, most of the accounts a single user has retweeted will share a common political bias. We derive the perceived political orientation of an account by the bias of the audience. The co-retweeted network visualizes the political beliefs of the public at large. For example, many news media accounts have been shown to have biased audiences (Golbeck and Hansen, 2011). Fox News and The Drudge Report, which are in the conservative group on our network, have a conservative audience, while National Public Radio and The New York Times have more liberal audiences and appear slightly to the left in our network. Our findings confirm the findings of (Golbeck and Hansen, 2011), however, our method is more flexible since it computes bias on specific events, not in general.

3.1 Network Polarity

A polarized network is composed of two or more groups of densely connected nodes that are linked through relatively few inter-group edges. We define as a network's *polarity degree*, *PD*, the ratio of sum of edges contained within each group over the total number of edges present in the network. Let E_i be the number of edges connecting nodes within group i and E_{ij} be the number of edges connecting nodes between groups i and j. Then, in a network containing k groups, its polarity degree is:

$$PD = \frac{\sum_{1 \le i \le k} E_i}{E}$$

Since $\sum_{1 \le i \le k} E_i = E - \sum_{1 \le i,j \le k} E_{ij}$, we have

$$PD = 1 - \frac{\sum_{1 \le i,j \le k} E_{ij}}{E}$$

Note that *PD* expresses polarization in a natural way: When the network is composed of few disconnected components (groups of nodes isolated from each other), $PD = 1$. On the other hand, when the network is complete, with each node representing a separate highly connected group, $PD = 0$. [9]

It is well documented that political life in America has become increasingly polarized (Pew Research Center for the People and the Press, 2012). Our analysis and visualization in Fig. 1 shows that this is indeed observed during the electoral debates, and computes its degree of polarization at 0.91. However, depending on the issue discussed, the polarization degree can rise more than 0.98, as it happens, for example in the online discussion with the hashtag #DoYourJobGOP (See Fig. 4). On the other hand, in Germany, a country where the online discourse was far less polarized

[9] We should note here a limitation of our definition of PD: Intuitively, it is easily justifiable for small k's, as is the case of parties represented in an election. If k is large, as in the case of family clans, the density of the overall graph should be taken into account.

than in the US and where coalitions between parties are frequent and expected, the co-retweeted network shows no comparable polarized divisions (Fig. 5). For the recent 2013 German elections co-retweeted network, the polarization degree was 0.48.

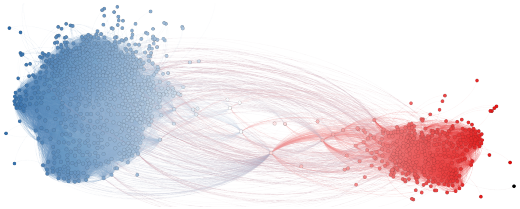


Figure 4: In a highly polarized online discussion in the US, marked by the hashtag DoYourJobGOP (indicating that the Republican party is not performing its duties in the Congress), the degree of polarization was close to 99%. A reason for this higher polarization is, no doubt, the fact that the hashtag was designed to divide as often happens in the so-called hashtag-wars.

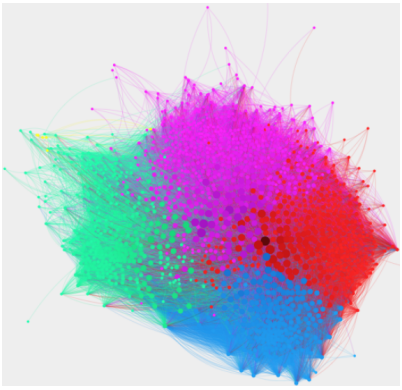


Figure 5: The co-retweeted network for the recent German elections shows far less polarization than during the recent US elections. The network's degree of polarization was about 48%.

3.2 Computing Account Polarity

Another related application of the co-retweeted network is to measure the polarity of political orientation for popular accounts according to how the audience sees their messages, not according to their own claims. The co-retweeted network is divided in two groups that exhibit liberal and conservative bias. Less polarized media actors, such as mainstream news media accounts, are retweeted by users of differing political bias and have co-retweeted links between accounts from both the liberal and conservative clusters in the network. In contrast, highly polarized accounts, such as political candidates, are less likely to be retweeted by users with opposing views and the majority of their links will be within their own cluster. This results in certain accounts being drawn into the center of the co-retweeted network, while others end up on the periphery of either the liberal or conservative groups. We utilize this observation to calculate the polarity degree of individual actors in the co-retweeted network.

Let us take the Oct. 16 co-retweeted network as an example for our description below. We define *an account's polarity value for a particular network* to be the normalized Euclidean distance of the account's node to a center of gravity of the network that typically falls between the two modules. (We normalize on a scale of -1 to 1, where the center 0 divides conservatives and liberals.) As a result, each account is assigned a polarity value based on their position in the network, as shown in Figure 6.

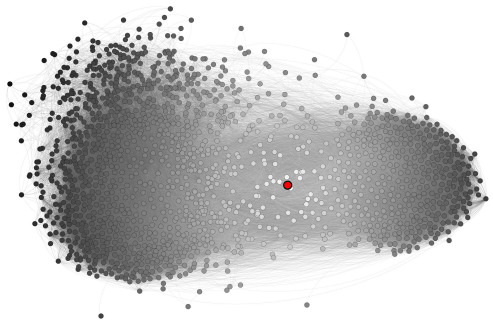


Figure 6: The co-retweeted network for October 16th, where nodes are darkened based on their polarity (darker nodes are more polarized). The large red dot represents the center.

From these polarity values, we can determine the polarity of individual accounts. For every user not included in this network who has retweeted at least two accounts in the co-retweeted network, we calculate their polarity as the weighted average of the polarities of the users they retweeted. A distribution of the 90,000 users for whom we calculated a polarity value shows that the Oct. 16 debate had a largely liberal leaning general audience (see Figure 7 (left)). However, if we look at the more active users that are likely more closely linked to the party's line than the rest, we see a more balanced audience: those who retweeted at least 10 accounts seem to be evenly distributed between the two parties (see Figure 7 (right)). We conclude that, even though both parties were actively trying to inject information in Twitter, the audience retweeted mostly the liberal messages. One could see this as an indication on which speaker in the debates was considered more popular by the online audience.

To compare the accuracy of the calculated polarities between the retweeted and co-retweeted networks, we performed the same computation in the

retweet network and we looked for some ground truth. We manually studied users who had words in their Twitter account description indicative of their political bias (ie, explicitly mentioning that they were democrat, republican, liberal, conservative). We randomly selected 77 such users with opposing polarities, and found that in 75 cases their polarity value accurately described them in the co-retweeted network. We take this as an indication that, even though both retweet and co-retweeted networks may be useful in determining political orientation, the co-retweeted network seems to be a bit more accurate in the determination.

The computation of the center was performed in an automatic calculation as follows. We first used the K-means algorithm to detect two clusters, and counted the cardinalities V_l and V_r of the left and right cluster. We also determined the center of gravity of each cluster c_l and c_r approximating them as circles. Finally, we computed the center of the whole system C as the *symmetric center of gravity* of the whole network, by placing it in distance relative to the ratio of the square root of the cardinalities. The x-coordinate of C is computed as:

$$x_C = x_l + \frac{\sqrt{V_r}}{\sqrt{V_r}+\sqrt{V_l}} \cdot (x_r - x_l)$$

where V_r and V_l are the cardinalities ("volumes") of the right and left clusters. The y-coordinate is computed similarly.

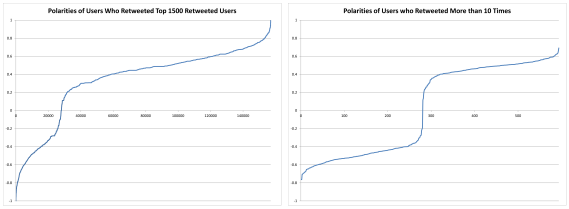


Figure 7: The distribution of polarity among (left) users who retweeted accounts in the co-retweeted network, and (right) those who retweeted accounts at least 10 times. It appears that, in this discussion while the audience was retweeting much more heavily the liberal tweets, the most active members of the audience were equally divided.

4 CONCLUSIONS

In this paper we introduced a novel network, the co-retweeted network. This is a meta-network that captures aggregated user behavior observed in Twitter during major real-time events. The core power of this network is that it can built on the collective intelligence of hundreds of thousands of humans who, with their retweeting decisions evaluate the major accounts tweeting these real-time events. We also presented two applications: the computation of the polarity of the event itself (as is represented by the network); and the computation of the polarity of the major accounts participating in the discussion (as perceived by the participating audience). Both of them are novel contributions in analyzing online political communication.

In the process we were able to answer the three questions that often are asked during political events: The computed polarity degree is able to answer the question of just *how polarized the political postings during an online event were*. In particular we found that the 2012 US presidential debates were highly polarized, much more than the discussion around the 2013 German elections, but not as much as the discussions around some online hashtag-wars designed to irritate the opponents and stir controversy.

The computation of a highly visible account's polarity, on the other hand, is able to answer the question on *whether major news organizations take sides in favor of one or the other candidate during the debates*. This computation is done per event, and not in general as other researchers have done in the past. This is important because news organizations do not always have a definite liberal or conservative approach for every issue. Their stance usually depends on the issue.

And finally we were able to answer the question on *whether the supporters of one debater were more vocal than the supporters of another*, by looking at the distribution of users' polarities retweeting liberal or conservative messages. We found that for the 2012 debates the answer is more nuanced than a simple yes or no: The distributions reveal that, while the core supporters may have worked equally hard to promote the message of their candidate, the audience was far be less divided in their support.

The concept of co-retweeting, inspired by bibliometrics co-citation and introduced in this paper, can incorporate important aspects of human computation power into the study and analysis of large online social media data. There are other applications one may find, such as its use in a recommendation system, however, describing it is outside the focus of this paper.

ACKNOWLEDGEMENTS

The authors would like to thank Andreas Jungherr and Pascal Jürgens for providing some of the data from the German elections and for insightful comments on an earlier version of the paper. This research was par-

tially supported by NSF grant CNS-1117693.

REFERENCES

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proc. of the 3rd Intl workshop on Link discovery*, pages 36–43.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10.

Cohen, R. and Ruths, D. (2013). Classifying political orientation on twitter: Its not easy! In *Proc. of 7th ICWSM*. AAAI.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *2011 3rd IEEE SocialCom*, pages 192–199.

Diakopoulos, N. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *CHI*, pages 1195–1198.

Finn, S. and Mustafaraj, E. (2013). Visualizing co-retweeting behavior for recommending relevant real-time content. In *Proc. of the 4th Intl Workshop on Modeling Social Media*, MSM '13, pages 4:1–4:2. ACM.

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the american public. *Annual Review of Political Science*, 11:563–588.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21:1129–1164.

Gephi (2010). ForceAtlas2, the new version of our home-brew layout. http://bit.ly/1deeWht.

Gerlitz, C. and Rieder, B. (2013). Mining one percent of twitter: Collections, baselines, sampling. *M/C Journal*, 16(2).

Golbeck, J. and Hansen, D. (2011). Computing political preference among twitter followers. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '11, pages 1105–1108, New York, NY, USA. ACM.

Golbeck, J. and Hansen, D. (2014). A method for computing political preference among twitter followers. *Social Networks*, 36:177–184.

Jungherr, A., Jürgens, P., and Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30(2):229–234.

Metaxas, P. T. and Mustafaraj, E. (2010). From obscurity to prominence in minutes: Political speech and real-time search. In *Proc. of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, 2010*, WebScience'10.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM*.

Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. T. (2011). Vocal minority versus silent majority: Discovering the opinions of the long tail. In *Proc. of 3rd IEEE SocialCom*, pages 103–110. IEEE.

Pew Research Center for the People and the Press (2012). Partisan polarization surges in bush, obama years. http://bit.ly/1d2HZK6.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, 16:101–127.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proc. of the 10th WWW*, pages 285–295. ACM.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 2nd IEEE SocialCom*, pages 177–184. IEEE.

von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom: a game for locating objects in images. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '06, pages 55–64, New York, NY, USA. ACM.

Wellesley Trails Group (2014). Retweets indicate agreement, endorsement, trust: A meta-analysis of published twitter research. Forthcoming.

Analysis of LinkedIn Privacy Settings
Are they Sufficient, Insufficient or Just Unknown?

Pilar Manzanares-Lopez, Juan Pedro Muñoz-Gea and Josemaria Malgosa-Sanahuja
Department of Information Technologies and Communications, Antiguo Cuartel de Antigones Campus Muralla del Mar s/n, Technical University of Cartagena, E-30202 Cartagena, Spain
{*pilar.manzanares, juanp.gea, josem.malgosa*}*@upct.es*

Keywords: Privacy, Professional Social Networking, LinkedIn.

Abstract: Internet-based applications give users an easy way to communicate with each other on a scale and rate unseen in traditional media. Among them, the professional social networking sites (with LinkedIn as one the most widespread platforms) offer a useful way to create and maintain a professional contact network. LinkedIn is also a self-promotion tool, where employees, industries and communities get in touch. In this scenario, it seems logical to consider privacy as a fundamental subject. Controlling who can see our data may avoid that our information reaches our boss when we are looking for a job, a competitor, or even former or present work colleagues with whom we have had some conflict. This work analyzes deeply the privacy settings offered by LinkedIn, and also analyzes the privacy concerns among the users, examining how these concerns correlate to the knowledge of the privacy settings and the adequacy of their use.

1 INTRODUCTION

Social media is defined as a group of Internet-based applications that built on the foundations of Web 2.0 and allow the creation and exchanges of user-generated content (Kaplan and Haenlein, 2010). They give users an easy-to-use way to communicate with each other on an unprecedented scale and at rates unseen in traditional media (Gundecha et al., 2011).

Among the variety of social media, social networking sites are platforms that enable users to interconnect by creating personal profiles (a representation of themselves), inviting friends and acquaintances to have access to those profiles, and sending e-mails and instant messages between each other. The largest social networking websites are Facebook and MySpace. Also belong to this type LinkedIn and Xing.

LinkedIn and Xing are focused towards business users who would like to maintain and extend their professional networks. Thus, profiles are usually strictly professional, including education and experience. In this scenario, it seems logical to consider that users pay more attention to the information they publish about themselves, as well as give more importance to the issues of privacy and security. Controlling who can see our data may avoid that our information reaches, for example, our boss when we are looking for a job, a competitor, or even former or present work colleagues with whom we have had some conflict.

The issues of privacy and security in the framework of online social networks have been studied by the research community (Gross and Acquisti, 2005)(Acquisti and Gross, 2006)(Krishnamurthy and Wills, 2008)(Fogel and Nehmad, 2009)(Rizk et al., 2009)(Stutzman and Kramer-Duffield, 2010)(Krasnova et al., 2010)(Liu and Gummadi, 2011)(Nayak and D'Souza, 2011)(Egelman et al., 2012)(Johnson et al., 2012)(Staddon et al., 2012)(Johnson, 2012)(Magazine, 2012). Some works analyze users' awareness, attitudes, and privacy concerns in social network sites from a generic point of view, although most of the studies are focused on Facebook.

To the best of our knowledge, no previous work has addressed this subject in a professional social networking website. We have chosen LinkedIn because of its overwhelming popularity, with more than 225 millions of users. Its main features from a privacy perspective are scarcely summarized in (Barrigar, 2009). In (Skeels and Grudin, 2009), a very interesting study of the benefits and utility of LinkedIn is exposed, but privacy aspects are not covered.

The main objective of this work is to analyze the privacy concerns among the users of LinkedIn (are LinkedIn users concerned about privacy?), examine how these concerns correlate to the knowledge of privacy settings (do LinkedIn users manage their privacy

settings correctly?) and observe the behavior the users take to protect their privacy (does the LinkedIn users behavior reflect their concerns?, do their privacy settings match their intentions?).

Social networking providers are often viewed as the source of privacy threats. Users fear that they can use their personal information for marketing purposes as well as share it with third parties. However, users may also face specific privacy-related dangers rooted in the public availability of their data.

The rest of the paper is organized as follows: section 2 describes the evolution of privacy subjects in the field of social networking websites. Section 3 describes briefly the website and analyzes deeply the offered privacy settings. The methodology used to collect the data to answer the formulated questions, and the analysis of the obtained data are described in Section 4. Finally, section 5 concludes the paper.

2 PRIVACY IN SOCIAL NETWORKING WEBSITES

Danah Boyd, one of the most influential women in technology, asserted that social media has prompted a radical shift from a "private by default, public through effort" world to "public by default, private through effort" one. In 2010, Facebook CEO M. Zuckerberg justified his company's decision to switch defaults to "everyone" with the logic that the youngest generation no longer cares about privacy. Similarly, Google CEO E. Schmidt claimed that if you have something that you don't want anyone to know, maybe you shouldn't be doing it. This thought was also shared by the founder of LinkedIn, when in 2010 affirmed that privacy is an issue for older people as younger people perceive the value of connection and transparency.

In this new scenario, privacy not only entails the personal action of deciding what we want to share and with whom, but also it is related to the knowledge, understanding and use made by people of the privacy settings offered by the social networking site.

One of the first studies about privacy revealed that FB users (basically university students) provided large amount of personal information, but privacy preferences were used by a small number of users (only the 0.06% of them) (Gross and Acquisti, 2005). A later study showed some changes. Although a relative majority of users was aware of the visibility of their profiles, still a significant minority wasn't (Acquisti and Gross, 2006).

In short time, Facebook stopped of being a social networking site limited to university students, and becomes the worldwide and intergenerational social networking website that we use today. In addition to a more wide variety of users, the increase of attention to privacy matters in the media caused that users increased the interest for the privacy subjects.

Although more users utilize the available privacy settings, researchers continue to identify inconsistencies between users' sharing goals and their privacy settings. (Liu and Gummadi, 2011) collected the privacy settings for all the uploaded content of 200 FB users, and reveled that 36% were shared with default privacy settings, a fraction higher than users who reported that this was the desired setting (20%). In addition, it outlined a more worrying conclusion. Even for contents for which the privacy settings have been modified, the modified privacy settings matched users expectations less than 40% of the time. This strongly suggests that users are having trouble to correctly configuring their privacy settings.

A study to evaluate how FB users react to limitations in the privacy controls was described in (Egelman et al., 2012). The results demonstrate that using the existing privacy controls many participants failed to complete the task, and that only ambiguity detection with actionable guidance improved participants' ability to complete the tasks.

(Johnson et al., 2012) measured the users' attitude toward privacy concerns on FB. They concluded that a great part of participants (86.2%) were either unconcerned with the threat of strangers viewing their content in FB, or they were able to mitigate those concerns through the use of global privacy settings. Thus, they belive that strangers are no longer the greatest thread. From their data, they conclude that threats from within users' friend networks are more concerning (because they are much less likely to be mitigated through the use of privacy settings).

In our opinion, this last conclusion is, perhaps, much more interesting in the field of LinkedIn, where our connections are not only friends or close colleagues, but also former and present bosses, colleagues and acquaintances in the work environment.

3 LinkedIn

The purpose and use of LinkedId will depend on each user, among its main objectives there is the self-promotion and maintenance of business and professional contacts. LinkedIn can be a great place to network with others in your career field and reconnect with some of your favorite former co-workers. It can be also considered as a free database of employees, industries and communities.

Controlling who can see your data or what infor-

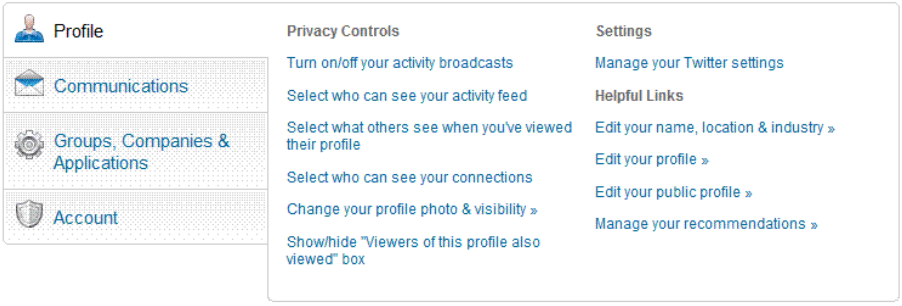


Figure 1: Settings page.

mation can be extracted from your actions and activities in this site, may avoid that your information reaches unwanted ears (for example, to your boss when you are looking for a new job) or falls into wrong hands with dangerous consequences for you.

3.1 A Brief Description of the LinkedIn Website

A key aspect to understand the LinkedIn privacy settings is to be clear about the difference between the user Homepage and the user Profile. When a user signs in on LinkedIn, the homepage is shown. There, a list of updates informs user about new connections of user's contacts, companies that are followed by user's contacts, modifications made by user's contacts to their profiles, about new groups joined by user's contacts, about information shared by user's contacts, about jobs, and of course about its own updates.

Unlike the user homepage, the user profile shows personal and professional information that it is wanted to be shared with others. All registered users can view a user profile, when they connect with him, search his name or click his name in an update, group, or other areas. The Profile can list your activities in LinkedIn (Activity Feed section), your education, past work history, and current and past projects (Background section), groups and associations (Groups sections), and more.

On the other hand, the user profile must be differentiated of the user public profile. The Public Profile appears when people search for a user using a public search engine like Google, Yahoo!, Bing, etc., that is, without being registered in LinkedIn.

Although professional, LinkedIn is a social networking site where users maintain a group of contacts. To facilitate the joining of new contacts to our network, LinkedIn offers two tools: the "People You May Know" list, and the "Who's Viewed Your Profile" list, which offers information about who has been looking at your profile in the last 90 days.

Finally, LinkedIn defines different levels of connections among users. 'Your connections' are 1st-degree connections, that is, people you're directly connected to because you have accepted their invitation to connect, or they have accepted your invitation. In addition, LinkedIn defines 'your network', that is made up of your 1st-degree, 2nd-degree (people who are connected to your 1st-degree connections), and 3rd-degree (people who are connected to your 2nd-degree connections) connections and fellow members of your LinkedIn Groups.

3.2 Privacy Settings

Users can change their account information and settings, through the settings page (https://www.linkedin.com/settings) or moving the control over the photo in the top right of the homepage and selecting "Privacy&Settings"[1].

Most of the privacy controls are located on the Profile tab (see figure 1). Below, the most important aspects of each one are described.

Turn on/off your Activity Broadcast. Some actions on LinkedIn trigger activity updates (also called activity broadcasts). By this option, the user decides whether or not to share these actions by means of activity updates. This control is on by default.

First of all, it is important to realize that this option only controls the generation of updates that will be displayed on the homepage of the others. Although this option is set to off, all those changes will be listed in the Activity Feed section of the user profile.

When a user accesses to this setting, LinkedIn advertises the following: *"Note: you may want to turn this option off if you are looking for a job and don't want your present employee to see that you're updating your profile"*. Although this note seems to be clear, a question is often repeated in LinkedIn discussion websites: if users turn off the activity broadcast,

[1]LinkedIn version at 24th. June, 2013.

make changes on their profiles and then turn on the activity broadcast, LinkedIn will provide feedback and report any changes made during the time the option was off? Clearly, the answer is negative.

Usually when a user is looking for a job, the profile is updated but also recommendations are posted. Regarding that, it is interesting to point out that an activity update is generated not only when a user makes a recommendation, but also when the user accepts a recommendation made by other LinkedIn member.

Finally, although the control description does not detail this aspect, joining a group (which gives tracks about our interests) or updating your photo generates an update although the control is turned off.

In our opinion, a possible improvement related to this control could be to offer a greater granularity. Thus, instead of activating or deactivating the generation of updates on the whole, users could select activate or deactivate the generation of updates associated with a certain activities in particular.

Select who can see your Activity Feed. A user can select who can see its Activity Feed section. The values are: 'Your connections' (the default value), 'Everyone', 'Your network', and 'Only you'.

Usually changing this control is associated to the previous setting. If you don't want to trigger updates to prevent your contacts being alerted about changes in your profile, you will not be interested in they can access to your profile and know, at a glance, your activities in the Activity feed.

Select what Others see when you've Viewed their Profile. This control helps you to protect your privacy when you are viewing other users profiles, selecting which information about your identity is shown. There are three options: 'Your name and headline' (Recommended and default option), 'Anonymous profile characteristics such as industry and title', and 'You will be totally anonymous'.

As can be seen, the first option is fully open, while the other two offer some degree of anonymity. The chosen value not only determines what information about us will get the owners of the profiles we have visited, but also determines what information we will get about the users who have visited our own profile.

If your LinkedIn account type is Basic and you decide to hide either partially or totally your identity (options 2 and 3), you will not get information about who has viewed your profile. The 'Who's viewed your profile' box will indicate the number of users who viewed your profile. But when you click on the link, the detailed information is not available. However, if you do not hide your identity, Profile Stats is available, offering the last 5 results of who's viewed your profile, the number of visits to your profile and the number of times you've appeared in search results.

On the other hand, the third option could be misleading. Although it is said 'you will be totally anonymous', the owner of the profile you have visited will see that 'Someone on LinkedIn' has viewed his profile, and this sentence is a link to a page called 'One of these people viewed your profile', where just a reduced set of 10 profiles are listed.

If your LinkedIn account is Premium, Profile Stats Pro is available independently of the selected degree of anonymity. Profile Stats Pro shows the full list of who's viewed your profile (you won't see additional information about a profile viewer if they've chosen to remain anonymous in their privacy settings), trends, total profile visits, keywords used to find your profile, number of times you've appeared in search results and industries of people viewing your profile.

In this case, the anonymity offered by the last two options is complete. Now, if option 3 is chosen ('totally anonymous'), the owner of the profile you have visited will see that 'LinkedIn Member' has viewed his profile, but now this sentence can not be clicked.

Select who Can See your Connections. Using this control, user can choose to show or not his entire list of connections. There are two options: 'your connections' (the default value), and 'only you'.

As advertised by a note, it is important to point out that although 'Only you' option is selected, people will always be able to see your shared connections.

On the other hand, users could think that choosing the 'only you' option, their list of connections will be safe. However, as described in (Staddon, 2009), that is not completely true. Using the LinkedIn Search engine, a user can discover, if not all, a large part of the list of connections of another user whose configuration aims to keep the list hidden. The procedure is briefly described below.

Consider that user A (the attacker) is a connection of user B (the target). B has set this privacy control to 'only you'. However, using the profile attributes configured by B, A will be able to discover part of B's contacts. For that, A must make note of the main attributes corresponding to the 'current', 'previous' and 'education' fields in user B profile. Next, A will use each of these attributes in the keyword field in the LinkedIn search tool. Then, for any returned user labeled '2nd', user A must click on the 'shared connections' link and check if B is listened there. If B is found, that user is a contact of B. This data mining process generates an incomplete list of B's contacts, which can be widen if the process is repeated using the attributes of the discovered contacts.

The cost of this procedure results in a repetitive task of introducing attributes and checking, one by

one, all the results. The higher the attacker's network of contacts, the longer the process. A large network will tend to lead to more hits for a given search, thus making the contacts of the target user harder to find.

Finally, the amount of connections discovered by the attacker is related to its type of LinkedIn account. If the attacker account is Basic, although the search returns more than 100 results, the attacker could only check the first 100 ones. If the account is Business, the number of results that can be checked increases to 300, to 500 if Business Plus, an to 700 if Pro Account.

Change your Profile Photo and Visibility. Users can upload a photo, which will be shown in their profile and also in the messages they send. A user can choose if the profile photo is visible to 'my connections', 'my network' or 'everyone' (the default value).

In our opinion, this control creates another point of confusion. Users who want that only their connections are able to see their pictures, would select this option, and trust on that nobody else can see the photo. However, this option only affects the LinkedIn profile, and not the Public profile. Although this control is used, the upload of the picture to the LinkedIn profile activates the showing of this photo in the public LinkedIn profile to everyone. If the user does not want to display the photo in the public profile, the corresponding control must be explicitly set.

Show/hide "Viewers of this Profile also Viewed" box. The "People also viewed" box, on the right side of the profile and homepage, shows some of the other profiles that viewers of your LinkedIn profile have also looked at. By this control, you can show (the default value) or hide this box. However, it is important to notice that disabling this option involves that your name won't show up in the "People Also Viewed" box on anyone else's profile.

LinkedIn advises that if you want to increase your visibility, keeping the module active can help potential employers, clients, recruiters and headhunters find your profile. By removing this module, you may decrease your visibility. However, the opposite can also occur. A potential employer could use the list shown in the "People also viewed" box, as a list of similar profiles that have been visited by other employers, that is, a list of competitors.

All the controls described up to now are located on the Profile tab and grouped under "Privacy Controls" header. However, another feature related to privacy is situated also on the Profile tab, but under "Useful Links" header (see figure 1).

Edit your Public Profile. Users can select 'Make my public profile visible to no one' or 'Make my public profile visible to everyone', the default option. In the last case, a user can choose what information will be viewable, selecting a combination of basic information (name, industry, location, recommendations), profile picture, headline, current positions, education, additional information and interested in.

Finally, another control related to privacy is located on the Groups, Companies&Applications tab, under the "Groups" header.

Turn On/off Notifications when joining Groups Users can set if they want to update to their network when they join a group. The default option is 'yes'.

The joining to new groups can give clues about our interests in new professional issues, for example when you are looking for a job. Thus, these notifications may be undesirable when our employers belong to our network. For this reason, LinkedIn advertises that *"You may want to turn this option off if you're looking for a job and want to be more private about which groups you join"*. However, it is important to remind that although the activity updates are disabled, a user always can visit our profile to check our information and updates.

4 METHODOLOGY AND RESULTS

The main objective of this work was to analyze the privacy concerns among the users of LinkedIn, examine how these concerns correlate to the knowledge of privacy settings and observe the behavior these users take to protect their privacy.

To collect data to answer these questions, we designed a questionnaire about the use of LinkedIn and in particular, privacy concerns, and which actions are taken by respondents to address these concerns[2].

The initial target population was composed of friends, former and current work colleagues, and acquaintances of the authors of this work. They received an invitation via e-mail to complete a web-based questionnaire. In addition, the invitation friendly requested their help in spreading the questionnaire, encouraging the receivers to forward the invitation to their contacts. They were informed about the objective of the study and were also informed that all information they provided would remain confidential.

A total number of 75 replies were collected during a period of 10 weeks in the middle of 2013. We expected a higher number of participants: we directly contacted with 500 people, each with an average of

[2]The questionnaire is available in https://docs.google.com/forms/d/1aTf1DlXakB8R9zfHS98LVtBWQQIYmz_M0uhCs6KBl1s/viewform

Table 1: Demographic characteristics of the sample and some data about the use of LinkedIn.

Question	Results
Q1. Gender	
Male	78.6%
Female	21.33%
Q2. Age	
under 21	0.00%
between 22 and 35	6.76%
between 36 and 50	86.49%
over 50	6.76%
Q3. Current professional status	
current or recent student	5.33%
worker	89.33%
unemployed	5.33%
Q5. Type of LinkedIn account	
Basic	97.30%
Premium	2,70%
Q6. How often do you visit LinkedIn?	
Several times per day	5,33%
Daily	22,67%
Occasionally	50.67%
Have profile, rarely use	16.0%
Have profile, never use	5.33%
Q11. Approximately, how many LinkedIn contacts do you have?	
less than 50	29.33%
between 50 and 100	22.67%
between 100 and 200	26.67%
more than 200	21.33%

100 friends, the collaboration in forwarding the questionnaire was asked, and other social networks were used. However the collected data allow us to extract very interesting conclusions, as described later.

Anticipating those conclusions, the number of participants is, in turn, a first and important finding. As it will be confirmed with the analysis of the results, anonymity in LinkedIn, the knowledge and use of privacy settings, and the consequences are unknown issues to users of LinkedIn, and even undervalued subjects. In our opinion, that is probably the reason the participation was low.

Table 1 shows demographic characteristics of the sample, and also general information about the use of LinkedIn. Most of the respondents are between 36-50 years old (86.49%). The number of users over 50, range that as the previous one belongs to the working age range, is much lower (6.76%). This large difference could be motivated by the technological gap between generations. The percentage of users within the range 22-35 years old, that corresponds to recent graduated users or in their first years of work, for which a tool like LinkedIn would be very useful for job search or promotion, is also low (6.76%).

Table 2: Statistics about non-professional information provided by users in their profiles.

Question	Results
Q7. Your profile name is...	
my real name	84.0%
my real name, but not complete	14.67%
a fake name	1.33%
Q8. Your profile photo is...	
No image	30.67%
I'm easily recognizable	68.00%
Semi-identifiable	1.33%
Group image	0.00%
Joke image	0.00%
Q10. Do you provide any fake information in your LinkedIn profile?	
Yes	5,33%
No	94.67%

One of the centerpiece of a social networking website's privacy posture is the privacy policy. In LinkedIn, the privacy policy is reachable by a link situated in the footer of the website. In spite of its importance in terms of privacy, a first result extracted from the data collected in the study is that only the 16% of respondents affirm that have read the LinkedIn privacy policy, and even 4% of participants affirm that they didn't know that the privacy policy exists.

Before evaluating in detail the privacy related results, we consider important to discuss the result from Q23 (*Do you feel that your information in well-protected by LinkedIn?*). Half of the respondents consider that their information is not well-protected. That is, the LinkedIn users still view the online social network providers as a source of privacy threats.

Table 2 shows statistics about non-professional information provided by respondents in their profiles. Almost 100% of them use their real name, and only one participant uses a fake name. Although it is a low percentage, 5.33% of respondents admit that they provide some fake information on the LinkedIn profile.

If we consult the existing literature in the writing of a traditional c.v., we can see that, depending on the geographical area, it is considered appropriate or not to include a picture. In Anglo-American contries, it is recommended not to include your picture, to avoid the possibility of being discriminated against for any reason associated with our appearance (sex, race, ...). However, in Spanish-speaking countries, a photo is generally expected at the top of the c.v.

This decision should be reconsidered in the new scenario created by the emergence and successful of professional social networking sites. Experts highlight that you are seven times more likely to have your

profile viewed if you have one (Kane, 2013). Another study reveals that recruiters spend an average of 19% of their time on your LinkedIn profile simply viewing your picture (Sullivan, 2013). Analyzing the results, the 30.67% of respondents indicated that they don't include a photo.

LinkedIn privacy settings not only allow users to decide if including or not a profile photo but also controlling its visibility. Q27 (*Do you think that it is positive that everyone can see your profile photo?*) measure the need of this privacy option according to the users perception. The 69.86% think that is positive, the 15.07 % think that is not positive and have limited the visibility, and the 15.07 % think that is not positive and they would like to limit the visibility.

However, in our opinion, it is also interesting to analyze the relationship between that perception and the inclusion of a photo in their profile. First of all it is interesting to realize that the 47.82% of participants who don't include a photo, consider positive its inclusion even if everyone can see it. Given that laptops, tablets and smartphones have integrated cameras, a possible improvement of LinkedIn could be the inclusion of a built-in tool to take and upload a photo when a user creates a LinkedIn account, instead of having to use an already existing one. This functionality would facilitate and encourage the use of a photo in the LinkedIn profile. The 21.74% of participants who do not include a photo indicate that they would like to limit the visibility, and 26.08% have decided to limit the visibility, just not using a photo. In regards to respondents who include a photo but do not consider positive that everybody could see it, slightly more than half do not know how to limit it.

The procedure to limit the visibility of the profile photo is extremely simple. The window, from which you upload a photo, asks you if your profile photo is visible to 'your connections', 'your network' or 'everyone' (the default option). However, in our opinion, this privacy setting could be confusing, since it is not clear if this control also affects to the public profile. Asked by this detail, our hypothesis is corroborated due to the fact that a high percentage of users do not know that the visibility of the photo in the public profile must be configured in other place (64.38%).

The existence of a LinkedIn profile and a public LinkedIn profile allows to configure in a different way the visibility of our information within and outside the LinkedIn network. Analyzing the obtained answers, it can be concluded that more than 80% of respondents do not feel the need to restrict the visibility of its public profile (Q12). For those that want to restrict the visibility of the public profile, only about half know how to do it. Q16(*Are you OK with (a) your classmate/collegues (b) your professor/boss (c) an employer (d)acquaintances (e)strangers looking at your LinkedIn profile?*) also asks about the need of profile visibility control, but now, in the LinkedIn network scope. From the results, Q16.(a) 97.33%, Q16.(b) 94.67%, Q16.(c) 98.67%, Q16.(d) 97.33%, Q16.(e) 90.67%, it can be observed that whatever the relationship with the visitor of the LinkedIn profile, less than 10% of respondents are bothered with the fact that the profile is visited. Comparing the results of Q12 and Q16, it can be inferred that the majority of LinkedIn users do not care that their profiles are completely visible, with even less concern if the visit is done from the own LinkedIn network.

Previous questions had as objective to evaluate the user attitude to the sharing of its profiles, from a global viewpoint. However, LinkedIn is not a simple website where to upload a digital version of our curriculum. In fact, our actions and activities, our contacts and a variety of data offer additional information about us. Only 16% of respondents affirm that they are concerned about the fact that people they know can obtain information about them from their LinkedIn activities. However, this percentage doubles if the observers are strangers.

LinkedIn allows to control the visibility of the Activity box, and also the generation of activity updates. Asked about both issues, approximately 80% of respondents affirm that they didn't know the former control, and approximately 70% affirm that they didn't know or were wrong with respect to the latter. Considering only those respondents that correctly defined the privacy setting "Turn on/off your activity broadcasts", the 27.27% of respondents indicated that they didn't know the existence of this option, 40.90% consider unnecessary its use, and only the 31.81% affirm that they have used this control.

Joining a group generates an update although the control is turned off. Asked about this detail, only one of them claimed to know this detail and how to modify the diffusion of this activity.

We have just analyzed the need and knowledge of the users about the visibility of their LinkedIn activities shown in the Activity Feed section. However, there are other actions that are not included in that section but offer information about us, and therefore, they are directly associated to privacy. On one hand, information generated from the visits made by users to other profiles, and on the other hand, information obtained from the list of connections of a user.

More than 90% of respondents want to know who has visited its profile. However, only the 61.33% do not care that the owners know that you have visited their profiles. That is, when the user is the agent of

the action, the privacy requirement increases. As previously mentioned, a privacy setting allows to control how our identity is shown. Asked about the knowledge of this control, it is very interesting to remark that, although more than half of respondents would prefer not to show its identity when consulting other profiles, a similar percentage are unaware of this option (concretely, the 68.92% do not know what information about them is shown when they visit a profile, the 34.32% have selected the option 'Name and Headline', the 5.41% have selected the 'partially anonymous' option, and the 1.35% have selected the 'totally anonymous' option). Finally, it is interesting to note that even those who know this control, largely are unaware of the result of using some kind of anonymity.

Next, we are going to analyze privacy concerns of users related to their connections. First, we would like to verify the need for the users of the privacy setting that allows to show or hide our list of connections. As in previous questions about other privacy controls, around the 60% of respondents do not consider negative that the list of connections is visible. Among the users that consider necessary to control the visibility, just 17.85% have hidden the list. An easy way to widen our list of connections is making use of the 'People you may know' tool. Due to the fact that we could appear as a possible contact of others, this tool affects our privacy. Observing the results (the 72% of respondants have added people using the 'People you may know' box), this tool is considered very useful and. In addition, a large majority (93.33%) of the respondents are not worried of appearing there.

Finally, the last question of the survey was on open question where the respondents were invited to indicate any privacy setting that they would like to establish and which is not currently offered in LinkedIn. Next, the contribution of participants are listed:

- *"I think that activities related to job search (headhunters, groups, following a company) should have a specific labelled or category which allows to easily avoid its broadcast."*
- *"I would be interesting to be able to manage the visibility of our information with a tool or concept such simple and intuitive as the Google 'circles'. I have created groups in LinkedIn, but its usage is not so easy."*
- *"I would like complete control over privacy."*
- *"Turn off activity broadcast means turn of ALL activity. Some of my activity still shows up for connections in the home feed."*
- *"I would like that LinkedIn were not so 'stool pigeon' as Facebook."*

However, some participants use this question to Zuckerberg and Schmidt:

- *"No, I don't care. I have nothing to hide. Being a free service, although with all their limitations, I think that it is wonderful."*
- *"The truth is that I have not been interested in the privacy settings of LinkedIn, so I don't know which are offered and which not."*
- *"I use this social networking site to make public the information I want to share with others. People who see and consult my information, the more the merrier. I think the LinkedIn is a tool which offers professional advertising that can become useful for me. I only upload what I want to be visible of my professional activity."*
- *"No. I think that it is useful to the professional promotion. You only must be careful with the information you upload, period."*

5 CONCLUSIONS

The first objective that we set in this work was to analyze the privacy settings offered by LinkedIn, mainly those which are oriented to allow users to control the visibility of the data and the actions they make on the professional social networking site. Through this analysis, we have been able to identify aspects that, in our opinion, are not sufficiently clear and could involve difficulties or even cause confusion.

On the other hand, we have been able to answer the questions posed in the objectives of this work. As a global conclusion, it can be affirmed that a considerable percentage of LinkedIn users are not worried about the privacy. Asked about the fact that their photos, profiles or their activities could be visible and known by colleagues, bosses, acquaintances or even strangers, a great majority do not express concern. However, approximately half of the participants consider that their information is not well-protected by LinkedIn. Users still continue identifying the social networking provider as the problematic agent which risks their privacy and security. However, users do not identify their own actions and choices (what and how they share their information) as a key aspect related to privacy and security. The above conclusion supports another one deduced from the conducted study: there is an important ignorance of the privacy settings offered by LinkedIn. Even considering those users who show interest and who have used some controls, there are constraints and consequences related to the used settings that are out of their knowledge.

Although it goes against the "public by default, private through effort" policy, which is assumed in this new online scenario, LI and other similar websites, should improve the way in which users perceive how their data are protected and how their privacy is guaranteed. After all, improving this perception will be translated into a greater trust and a more intensive and wider use of the social networking. In our opinion, a solution could be to change the privacy settings interface, which is composed by multiple items located separately, and offer a simpler, explanatory and self-guided tool to configure the privacy settings. This interface should be shown to users when a LinkedIn account is created, and later, periodically.

ACKNOWLEDGEMENTS

This research has been supported by project grant TEC2010-21405-C02-02/TCM (CALM). It is also developed in the framework of "Programa de Ayudas a Grupos de Excelencia de la Region de Murcia, de la Fundacion Seneca, Agencia de Ciencia y Tecnologia de la RM (Plan Regional de Ciencia y Tecnologia 2007/2010)".

REFERENCES

Acquisti, A. and Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the fb. *Privacy Enhancing Technologies*, LNCS 4258:36–58.

Barrigar, J. (2009). Social network site privacy. a comparative analysis of six sites. Commissioned by the Office of the Privacy Commissioner of Canada.

Egelman, S., Oates, A., and Krishnamurthi, S. (2012). Oops, i did it again: Mitigating repeated access control errors on facebook. In *Proc. of CHI'12*. ACM.

Fogel, J. and Nehmad, E. (2009). Internet social network communities: Risk taking, trust, and privacy concers. *Computers in Human Behavior*, 25(1):153–160.

Gross, R. and Acquisti, A. (2005). Information revelation and privacy in online social networks (the facebook case). In *Proc. of WPES'05*. ACM.

Gundecha, P., Barbier, G., and Liu, H. (2011). Exploiting vulnerability to secure user privacy on a social networking site. In *Proc. of ACM SIGKDD'11*. ACM.

Johnson, M., Egelman, S., and Bellovin, S. M. (2012). Facebook and privacy: It's complicated. In *Proc. of SOUPS'12*. ACM.

Johnson, M. L. (2012). *Toward Usable Access Control for End-users: A Case Study of Facebook Privacy Settings*. PhD thesis, Columbia University.

Kane, L. (2013). 8 mistakes you should never make on linkedin. http://www.forbes.com/sites/learnvest/2013/03/04/8-mistakes-you-should-never-make-on-linkedin/.

Kaplan, A. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53:59–68.

Krasnova, H., Kolesnikova, E., and Gunther, O. (2010). Leveraging trust and privacy concerns in online social networks: an empirical study. In *Proc. of ECIS'10*.

Krishnamurthy, B. and Wills, C. (2008). Characterizing privacy in online social networks. In *WOSN, 1st Workshop on Online Social Networks*. ACM.

Liu, Y. and Gummadi, K. (2011). Analyzing facebook privacy settings: User expectations vs. reality. In *Proc. of IMC'11*. ACM.

Magazine, C. R. (2012). Facebook and your privacy: Who sees the data you share on the biggest social network? http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm.

Nayak, A. and D'Souza, R. M. (2011). Who do you trust? information sharing, privacy concerns and trust in an online social network. *Explorations Journal*.

Rizk, R., Marx, D., Schrepfer, M., Zimmermann, J., and Gunther, O. (2009). Media coverage of online social network privacy issues in germany - a thematic analysis. In *Proc. of AMCIS'09*.

Skeels, M. and Grudin, J. (2009). When social networks cross boundaries: A case study of workplace use of facebook and linkedin. In *Proc. of GROUP'09*. ACM.

Staddon, J. (2009). Finding hidden connections on linkedin an argument for more pragmatic social network privacy. In *Proc. of AISec'09*. ACM.

Staddon, J., Huffaker, D., and Brown, L. (2012). Are privacy concerns a turn-off? engagement and privacy in social networks. In *Proc. of SOUPS'12*. ACM.

Stutzman, F. and Kramer-Duffield, J. (2010). Friends only: Examining a privacy-enhancing behavior in facebook. In *Proc. of CHI'10*. ACM.

Sullivan, J. (2013). Why you can?t get a job recruiting explained by the numbers.

Using PageRank for Detecting the Attraction between Participants and Topics in a Conversation

Costin Chiru, Traian Rebedea and Adriana Erbaru
University Politehnica of Bucharest, Department of Computer Science and Engineering,
313 Splaiul Independetei, Bucharest, Romania
{costin.chiru, traian.rebedea}@cs.pub.ro, adriana.erbaru@cti.pub.ro

Keywords: CSCL, Natural Language Processing, Participant Assessment, PageRank, Online Conversations.

Abstract: In this paper we present a novel approach that uses the well-known PageRank algorithm for assessing multi-threaded chat conversations. As online conversations can be modelled as directed graphs, we have investigated a method for allowing a real-time analysis of the conversation using PageRank by computing the ranks of the utterances based on the explicit and implicit links available in the discussion. This model has been also extended to offer a method for computing connections between the debated topics and the chat participants and between each of the debated topics in the conversation, called the participant-topic and the topic-topic attraction. The results presented in this paper are promising, but also reflect several important differences between the existent offline analysis tools for chats and the PageRank method.

1 INTRODUCTION

Chat conversations (instant messaging) represent nowadays one of the most popular methods of exchanging ideas online. The easiness in learning how to use chats and the high efficiency in transferring the information, promoted chats as one of the favourite environments for Computer Supported Collaborative Learning (CSCL) tasks requiring online and synchronous textual interactions among participants (Stahl, 2006; Stahl, 2009). Due to this fact, it has been largely adopted in CSCL activities and it has been enhanced with functionalities specific to these tasks such as the explicit referencing mechanism and the whiteboard facility present in ConcertChat (Muhlpfordt and Wessner, 2005). Still, in spite of its popularity and of the huge quantity of data that is exchanged through chats, there are very few application aimed at analyzing this type of content (Chiru et. al, 2011; Rebedea et. al, 2011). More than that, the existing applications are built starting from a semantic analysis (Chiru et. al, 2011; Rebedea et. al, 2011) but the analysis takes far too much time to be used as a real-time process and can only be applied offline, at the end of the conversation. Therefore, we have been searching for a different method to analyze these conversations faster and got influenced by the algorithms which are used by search engines that have to analyze huge quantities of data in a very short time. Thus, we reached the conclusion that if the PageRank algorithm (Page et. al, 1998) could be adapted for chats, this method could be applied online (displaying the results of the processing as the conversation unfolds) and interactive (to signal what threads should be debated more and involving people who contributed less on specific threads), this way improving the learning process and enhancing the participants' innovation.

To achieve this, we started from PolyCAFe (Rebedea et. al, 2011), a system that is using innovative methods for analysing CSCL chat transcripts, helping both computer-assisted learning and the tutors in evaluating the discussions. This system analyzes chat logs using Natural Language Processing (NLP), Latent Semantic Analysis (LSA) and Social Network Analysis (SNA) techniques in order to identify the most important utterances from the conversation (in terms of their content and of the participants' involvement in the discussion).

Therefore, this paper presents an extension of PolyCAFe's functionality, trying to enhance it with the ability to analyze CSCL chat sessions in real-time. The first step of our analysis consists of detecting the important utterances from the conversation using the PageRank algorithm. Once these utterances are identified, we use the PageRank

algorithm for detecting what is the attraction of each participant towards the debated topics and, at the same time, what is the probability of a topic to follow another topic within the conversation.

Most of the existing approaches for analyzing text using social network analysis (SNA) tools are oriented towards systems that own explicit referencing tools, such as forums or blogs. The reason behind this orientation is the ease in constructing the participant social network based on the order in which the messages are sent and on the recipient of the message. Still, there are a few systems that intended to apply SNA tools to chat conversations. One such tool was built by Sundararajan (2010) for analyzing the content published by the participants to 8 different courses in order to observe how the respect and influence earned by each participant influences their efforts to "collaborate, learn new and conceptual knowledge" and their satisfaction regarding the courses outcome. Unfortunately, the author does not mention whether this analysis is done manually or automatically. Moreover, a regular SNA method is used for evaluating the participants from the perspective of their centrality, betweenness, in-degree, out-degree, etc. in the network, which represent only quantitative data. On the other hand, we are rather interested in what the participants communicate (what are the topics they know or they are interested on) and in the interaction patterns between different concepts that are debated in the conversation, which is part of a qualitative evaluation of the participants, topics and the conversation as a whole.

A more similar approach was undertaken by Tuulos and Tirri (2004). The authors present a semi-supervised system that uses a combination of topic modelling and SNA to improve the information retrieval from chat conversations. For their analysis, they have used conversations taken from SearchIRC.com which allowed them to use simple heuristics in order to identify to whom each utterance is addressed (and therefore to build the social network). For this participant network the in-degree, out-degree and PageRank of each participant are determined. After that, the authors use some existing conversations to detect the probabilities of words to appear in conversations about different topics, so that when they analyze new conversations to be able to use these probabilities. Finally, they evaluate the use of each of the SNA technique in improving the information retrieval, considering as baseline the results provided by the topic modelling. Still, this approach gives them two advantages: first of all they know both how many and what topics should be present in the conversation (therefore knowing what represents off-topic and being able to discard that part); secondly, they have chosen the topics from different topics (Bible, C++, Philosophy, Physics, Politics, Win2000) thus simplifying the task of identifying to what topic a given concept corresponds. In our approach neither of these facts can be exploited: since our system does not have a learning phase, it gives the possibility to analyze texts debating about any topics, without being limited to the ones that were learnt (thus providing generality in use). At the same time, it can be used to distinguish between concepts that are from the same or similar conceptual area. The examples presented in this paper contain concepts from a single domain (Human Computer Interaction) especially to prove that the approach works even at this level, without requiring that different topics to be debated in the same conversation.

The paper continues with a short overview of the PageRank algorithm. Then, we present the application that has been developed and several results that have been obtained by employing the PageRank method adapted for CSCL chats. The paper ends with an analysis of these results and with our conclusions regarding the improvement of the results' quality.

2 OVERVIEW OF THE PAGERANK ALGORITHM

Because previous researches have modelled an online conversation as a graph with implicit and explicit links between utterances (Rebedea et. al, 2011), we have started to consider that the PageRank algorithm (Page et. al, 1998) may be a candidate for the conversation graph analysis. PageRank is an algorithm that was initially designed for the analysis of a set of web pages in order to extract the relative importance of each page from the considered set of web pages (Page et. al, 1998). The algorithm expresses the probability that a web surfer will be able to "find" the considered page within a limited number of steps (clicking on the links from one page to another). It is a customization of a "random walk" in a graph, which in turn is modelled as a Markov chain in which the states are pages, and the transitions, which are all equally probable, are the links between pages.

The formal definition given in the initial paper describing PageRank (Page et. al, 1998) was: if u is a web page; *Fu (forward links)*, the pages referred

by *u; Nu* $= |Fu|$ the number of forward links; *Bu (backward links)* the ones that refer *u*, *c* a normalization constant and E(*u*) a source of rank to make up for the rank sinks (such as cycles) with no out-edges, than the value of the rank R(*u*) can be computed using:

$$R(u) = c * \sum_{v \in B_u} \frac{R(v)}{N_v} + c * E(u) \quad (1)$$

In order to compute the vector R(*u*), one starts from the square matrix (we'll call it *A*) having the web pages on the rows and columns and $A[u, v] = 1/Nu$ if there is a link from page u to page v or 0 otherwise. If *R* is the a vector of scores over the web pages, then we can write $R = c(AR + E)$, which can be re-written as $R = c(A + E \times 1)R$ because the values of *PR* are normalized and therefore $\|R\|_1 = 1$. That means that *R* is the eigenvector of $A + E \times 1$ and the method should also try to maximize the value of *c* (Page et. al, 1998). Subsequent research showed that the optimal value for *c* should be 0.85 (Brin and Page, 1998).

The value of *R* can be obtained in an iterative manner, starting from a vector of values over the web pages (*S*) that can have any values (could be the vector *E(u)*), using the following iterative algorithm (Page et. al, 1998):

$$\begin{aligned} &R_0 \leftarrow S \\ &loop: \\ &\quad R_{i+1} \leftarrow A_i \\ &\quad\quad d \leftarrow \|R_i\|_1 - \|R_{i+1}\|_1 \\ &\quad R_{i+1} \leftarrow R_{i+1} + dE \\ &\quad\quad \delta \leftarrow \|R_{i+1} - R_i\|_1 \\ &while\ \delta > \epsilon \end{aligned} \quad (2)$$

, where d is a factor for increasing the convergence rate and for maintaining $\|R\|_1 = 1$, while $\epsilon = 10^{-3}$.

A web page will have a high PageRank if the sum of the webpages' PageRank that refers it is large. This property covers two possible cases: when a page has many other pages referring it, or when it is referred by pages with high PageRank.

The PageRank algorithm has proved to be suitable not only for Google's rank of web pages, but also for other tasks in various domains: replacing the ISI factor with a new formula based on the PageRank Algorithm (Bollen, Rodriguez and Van de Sompel, 2006), ranking academic doctoral programs based on their records of placing their graduates in faculty positions (Schmidt and Chingos, 2007), predicting how many people (pedestrians or vehicles) come to the individual spaces or streets (Jiang, 2006), performing Word Sense Disambiguation (Navigli and Lapata, 2010), etc. Thus, we hoped that it could also work for chat analysis especially as a conversation can be seen as a graph of links between utterances where discourse flows in a similar manner to the importance of the web pages.

In order to use this approach, we considered that each utterance from the chat conversation represents a different document and in order to simulate the forward and backward links, we used explicit and implicit links (details will be provided in the next section). Thus, we managed to develop a method for very fast identification of the important utterances from a chat, of the major threads of discussion, and of the participants' attraction towards these threads.

3 PAGERANK FOR AUTOMATIC ASSESSMENT OF CSCL CHATS

As we have already mentioned, our application starts from PolyCAFe project and uses some of its features:

- Detection of the areas of high collaborative discourse from a chat;
- Evaluation of the collaboration of each participant in the discussion (based on multiple criteria);
- Graphical representation of the results.

3.1 Pagerank for Conversation Analysis

We consider that PageRank is appropriate for chat evaluation as this problem is very similar to the original problem for which it was initially designed. The sparsity in the chat is (in our opinion) similar to the sparsity of relevant content from the web. Therefore, we consider that the probability of a person to "land" on a specific page from the web is similar to the probability of a participant to reply to a given utterance, while the links between different pages are well simulated by the semantic connections represented by the repetitions of the same word and by the explicit "reply to" links. Practically, this probability of a participant to reply to a given utterance can be considered the "rank of an utterance" (and it highlights the importance of that specific utterance in the conversation).

The first step in applying the PageRank Algorithm was the identification of the links that exist between chat utterances. Since the pre-processing part of our application was borrowed from PolyCAFe, we also kept the input format of the

chats, which allowed the existence of explicit links (references provided by the chat participants to specify to which previous utterance their answer is addressed).

Besides the explicit links, one can also encounter the situation when two or more utterances contain concepts that are strongly related and therefore their authors consider that there is not necessary to provide an explicit link. We considered that this situation is a special case of connection (an implicit link) and tried to identify it in order to augment the number of explicit connections (that was insufficient for our purposes). Therefore, we considered that words repetitions (Chiru et. al, 2011) are example of such links. If a term appears in an utterance, all lines that follow and contain that term are considered implicit links to the initial utterance. Given the nature of the algorithm, the two types of links that we consider (explicit and implicit) have equal weight.

Once we detect all these links, we build utterance chains (which can be interpreted as discussion threads since they debate the same concepts) starting from these links using the DFS (Depth First Search) Algorithm, thus finding all the existing separate chains. They are needed for determining the attraction between two different threads (topics).

The steps that should be followed in order to determine the threads are:

1. Identify all the utterances that are not referenced (neither by explicit nor by implicit links) – these utterances are probably off-topics and therefore they are ignored;
2. All the remaining utterances are considered to be roots for the DFS Algorithm;
3. From each of these utterances (considered in the order they appear in chat) we start a function (implementing DFS) to detect the threads that can be built starting from that utterance;
4. Each function will return a thread of utterances.

The next step is to create the transition matrix corresponding to the chat utterances by considering the links identified between them. The explicit and implicit links between two utterances will provide a value of 1 in the matrix, while the remaining elements are set to 0 (meaning that there is no connection between the corresponding two utterances). Once this matrix is built, it needs to be normalized with respect to the sum of the elements from each column.

Finally, the values of the PageRank algorithm for the given matrix are obtained using the power method implementation provided by the JAMA library - A Java Matrix Package (Hicklin et. al, n.d.).

The operations made for the detection of the eigenvalues and the eigenvectors are:

1. Apply the *eig* method, which decomposes the matrix in two other matrices: a matrix *D* containing the eigenvalues and a matrix *V* containing the eigenvectors;
2. The maximum (dominant) eigenvalue from the diagonal matrix *D* is determined and its index is stored;
3. The dominant vector is the column from the matrix *V* having the index identified in the previous step;
4. The values from this vector are normalized with respect to the sum of its elements;
5. The final values (the PageRank) are the values obtained for the normalized eigenvector *vd* (*utterance[i].rank = vd[i]*).

Once these values are determined, one can evaluate the participant-topic attraction and the topic-topic attraction as described in the following sections.

3.2 Participant-topic Attraction

The *participant-topic attraction* defines the participants' drive to get involved in the discussion of a given concept therefore proving its interest or knowledge related to that concept. To determine this factor, we have used the values of the participant's utterances containing the words that define the considered topic.

If p represents a participant and t a topic, then the attraction between p and t is given by the following formula:

$$a(p,t) = \sum_{\substack{up \in Utt \\ up \in p}} rank(up) * freq(t, up) \quad (3)$$

In order to highlight this method, we provide an example that proves how the above formula works. For this, we have made the simplifying assumption that all the utterances belong to the same participant.

Utt:	**< debated topics >**				**utt value**
u1:	t	t	y	x	0.5 = rank(u1)
u2:	t	y	z	x	0.3 = rank(u2)
u3:	t	t	t	z	0.2 = rank(u3)

Using formula (1), the following results will be returned (2):

$$a(p,t) = 2 * rank(u1) + rank(u2) \quad (4)$$

$$+3 * rank(u3)$$
$$a(p,x) = rank(u1) + rank(u2)$$
$$a(p,y) = rank(u1) + rank(u2)$$
$$a(p,z) = rank(u2) + rank(u3)$$

In the end, all these values are normalized.

3.3 Topic-Topic Attraction

The *topic-topic attraction* defines the probability of having a specific topic following another topic in the flow of a conversation. To determine it, we use the utterance chains taking into account both the frequency of each topic and the case when they co-occur in the same utterance (topics are very closely related) or occur separately (more loosely).

Therefore, we extract the threads corresponding to the two topics and build a matrix for each chain. This matrix reflects the debating of those topics in the utterances and their corresponding values within that chain.

The relationship between the topics and the utterances is reflected by the matrix that is built as follows:

1. We build the *chain – topic matrix* (*ct*), $ct[i][j]$ represents the value of the j topic in the i utterance where
   a. $ct[i][j] = 0$ if the i topic is not debated in the utterance j;
   b. $ct[i][j] = \frac{rank(utt\_j)}{\#topics\ in\ utt\_j}$ if the i-th topic is debated in the utterance j.
2. After filling in the matrix, we apply formula (5) for each topic t_i and t_j.

$$\frac{\sum_{u \in utt\_chain} ct[u][t_i] + \sum_{u \in utt\_chain} ct[u][t_j]}{\sum_{\substack{u \in utt\_chain \\ t \in t_i \mid t_j}} ct[u][t]} \quad (5)$$

Below we present an example of matrix (6) for determining the *topic – topic attraction* for the following chain: u5 → u4 → u3 → u2 → u1

utt:	**<debated topics >**		
u1	t1	t2	t3
u2 :	t1		t5
u3 :	t2		t4
u4 :	t1		t3
u5 :	t1		t2

$$ct = \begin{pmatrix} rank(u1)/3 & rank(u1)/3 & rank(u1)/3 & 0 & 0 \\ rank(u2)/2 & 0 & 0 & 0 & rank(u2)/2 \\ 0 & rank(u3)/2 & 0 & rank(u3)/2 & 0 \\ rank(u4)/2 & 0 & rank(u4)/2 & 0 & 0 \\ rank(u5)/2 & rank(u5)/2 & 0 & 0 & 0 \end{pmatrix} \quad (6)$$

Then, the attraction between topic t1 and t2 topic is given by (5) by applying the formula (3), with the matrix from (4), obtaining formula (7).

$$a(t1,t2) = \frac{rank(u1) * 2/3 + rank(u2)/2 + rank(u4)/2 + rank(u5)}{rank(u2)/2 + rank(u3)/2} \quad (7)$$

3.4 User Interface

The user interface allows the input file selection, and afterwards the content of this file is analyzed and the results are displayed in tabular form (see Fig. 1).

The left part of the GUI presents the values for participant – topic attraction for the selected participant, while the right part gives the values of the topic – topic attraction for the selected topic.

Besides the values obtained for the participant – topic attraction and the topic-topic attraction, the application also outputs the most important utterances from the chats computed using their PageRank.

The results proved to be much stricter comparing to the results obtained using PolyCafe system or provided by the human reviewers (Gold-Standard). This is due to the fact that only very few utterances have a PageRank greater than 0.

In order to provide an example, we present a part of the utterances evaluated as being important by PageRank algorithm. We will use the same chat for which we presented the examples from the participant-topic and topic-topic attraction examples. The automatic analysis performed with PolyCAFe has identified 132 important utterances (out of 430) as important. From these, the PageRank algorithm also identified 17 utterances as (see Table 1). The values for the ranks computed by PageRank may seem pretty low, but this is what usually happens when applying the algorithm on any graph.

Besides these 17 utterances that were considered important by both PolyCAFe and the PageRank method presented in this paper, the latter has identified another 28 turns that were not considered important by the former. In order to account for these utterances, we analysed PolyCAFe's results in order to discover a possible explanation. At a careful analysis, we observed that the 28 extra utterances were marked by PolyCAFe as being continuations of other utterances. Therefore, it is possible that PolyCAFe did not consider these utterances to be

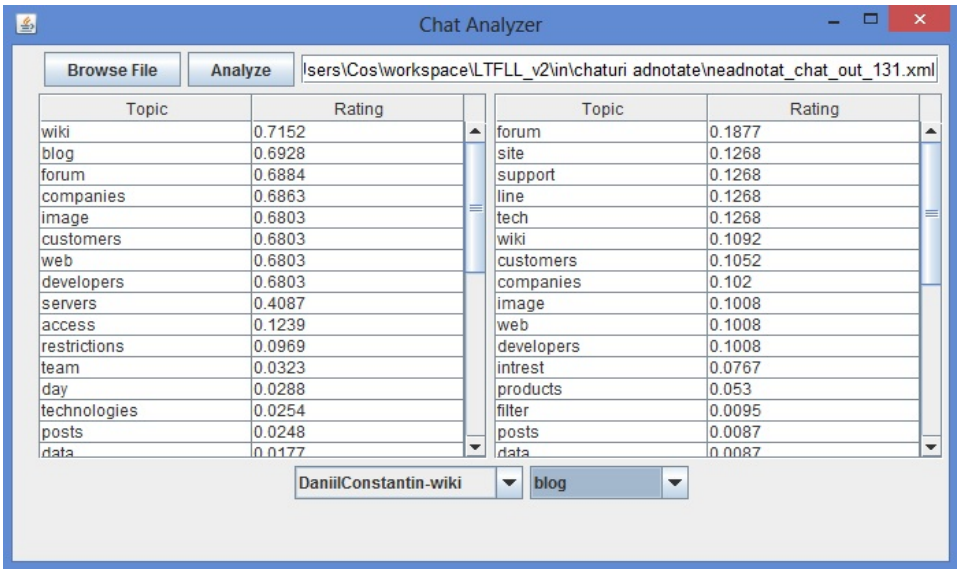


Figure 1: Application Graphical User Interface.

Table 1: The utterances identified by our algorithm that receive a grade higher than 8 by PolyCAFe.

Utt. No	PageRank score	PolyCAFe score	Utterance Content
169	0.002	10.07	yes, they have wikis that are publicly available, with public information, for the everyday user that takes an intrest in that company's products
167	0.002	10.01	all major companies have wikis for their technologies. most people like to search wikis cause they provide accurate and easy to access information. Also, that way our database servers won't be so used
404	0.202	10.01	Indeed. Our companies image will grow if we have a forum, a blog, a wiki and a cool web-site that customers or developers can use
310	0.004	9.93	the only problem that still remains is that we need someone to check wiki articles, blog and forums posts so that classified data does not accidentally reach a "public" area
348	0.004	9.28	A svn is an open-source revision control system. Users can work on a version of the application code and commit it. If two users are working on the same thing when they commit a merge is made with the 2 versions
308	0.001	9.03	we can use a person or a team of people to handle the wiki posts, forums posts, wave documents and all the other important stuff
331	0.002	8.93	I mean everyone of our employees knows how to use a wiki, forum blog and chat, and google wave has a extremely friendly interface
314	0.003	8.85	but how can you use a filter for a forum?
399	0.009	8.79	well i think chat is important for our employees, it helps them talk and colaborate, spare time by not meeting in conferences that much, and be on track with all theit colleagues are doing
312	0.002	8.78	We can use filters for that firewalls. That can save some money
388	0.047	8.45	not necesarly computers,you can change acounts
333	0.001	8.4	evrybody can use a chat , forum , bog or something like that
423	0.015	8.32	good night everyone, and thank you for your collaboration:)
341	0.004	8.21	We can also use a SVN for our code. What do you think of this?
376	0.042	8.12	this could make them loose time...
373	0.029	8.08	they will use another machine. the restrictions will be only for certain computers
387	0.025	8.07	for every function you don't remember in a programming language, you will have to move to other computer to find out... but it's ok ... it wouldn't be a big problem i guess

important, since the same ideas were present in the previous utterances, but PageRank, through its nature, favours this kind of utterances since being identified as continuations it means that they have links from other utterances that were considered important in the past and therefore they receive a part of these utterances' rank.

For a better evaluation of our method, we asked 30 students from the Human-Computer Interaction class to annotate 3 different chat conversations with the most important utterances. Thus each of the three chat conversations was evaluated by 10 different students. We computed the inter-rater agreement using Fleiss' Kappa for m raters and we have obtained the values for Kappa 0.133 for the first chat, 0.142 for the second and 0.177 for the third one, while the p-values were always 0.000. These results show how difficult this task is even for humans. When we computed the results obtained by our method with the gold standard results provided by the annotators, we obtained the values of kappa 0.085 for the first chat, 0.0882 for the second and 0.0894 for the last one. These results are below those of the raters, but we have remarked that if we discard the last 15% of the utterances in all chat conversations, where the PageRank accumulated too much, the results are much better: 0.131 for the first chat, 0.128 for the second and 0.173 for the last chat conversation. These results are closer to the inter-rater agreement and highlight that we should add a decaying factor for utterances that are closer to the end of the discussion (as they have fewer out-going links and thus the rank tends to accumulate in them).

4 INTERPRETATION OF RESULTS

There are several important observations that can be drawn up based on the results that we have obtained and analyzed. First of all, the computed ranks for the utterances are rarely different from 0, this fact being generated by several reasons:

- Most of the chats contain very few explicit links (they seem to be ignored quite often by the chat participants) – we have observed a direct dependence between the number of explicit links from the chats and the number of utterances having non-zero values after applying the PageRank algorithm.
- The PageRank algorithm determines the utterances' ranks as a random walk in the graph of utterances. The significance of these values is the probability to get to a certain utterance after a number of steps that goes to infinity. Therefore, once the algorithm gets to a (relatively small) set of utterances (lines / columns from the transition matrix), it will be very difficult to get out of that set (in the context of random walk) and so the remaining values will tend to be 0.
- The PageRank Algorithm is designed for the web, where a lot of links exist between different resources (therefore creating large chains of links, most of the times having a lot of cycles), while in the chats the utterance chains are usually short and rarely having such cycles. Besides, the topics repetitions might not be synchronized with the explicit references, therefore not leading to cycles.
- In the current version of the system, we proposed equal importance for the explicit and implicit links, which can lead to determining a value too high / low for some utterances, depending on the number of words used in that utterance and in the one to which it is linked.

Secondly, most utterances having values greater than 0 are positioned at the end of the discussion. This happens due to lack of explicit links, and therefore those utterances accumulate a very high score due to topics repetition, which propagates through the chat from the beginning until the end. A solution is needed to link these utterances to other ones in the chat.

Finally, there are some utterances considered significant by the PageRank algorithm, but not by other algorithms or by human evaluators. The reason is the same as for the previous observation: some utterances (that are not very significant in terms of conversation) may receive high values because of the rank accumulation over time from other utterances that contain the same topics.

There are a couple of solutions that could be tried in order to alleviate the presented problems. A first solution might be the detection of dialog acts and adjacency pairs, since the main problem of the proposed method is the lack of explicit links. This way, one can detect the dialog acts that are present in the chat (question - answer, agreement – disagreement, greetings and so on) very quickly and to use these links as additional explicit links. Another possibility is to use LSA or lexical chains to find out more semantic connections between utterances.

Another solution to avoid reaching too many zeroes for the computed ranks is to use the Iterative Method instead of the Power Method for computing the PageRank values. This way, one is not

constrained to apply the algorithm until convergence (after an infinite number of steps), but can stop after a limited number of steps, so that fewer utterances reach a zero-value influence.

Finally, in order to be able to discriminate between the importance of explicit and implicit links, one can use a generalized algorithm based on Markov chains having different values for different link types (explicit or implicit links).

5 CONCLUSIONS

To sum up, our current adaptation of the PageRank algorithm for online conversations (using only the explicit and implicit links given by the topics repetitions) is not powerful enough to provide results that have the desired accuracy compared with other solutions that analyse the discussions offline. The main explanation is that there are not enough explicit links added by the participants during the discussion and using only repetitions for detecting implicit links does not build a graph that is dense enough. However, the method is much faster and it can be used online and in real-time for the dynamic evaluation of multi-threaded discussions involving multiple participants.

Moreover, in our opinion the assumptions made in this paper are novel for the analysis of online discussions and they have not been used to assess the importance/rank of an utterance in an online discussion although PageRank follows from previous work in citation analysis (where the links between papers are made explicit by authors). The preliminary results also support the use of PageRank to compute the most important utterances in a multi-party online conversation, but several improvements of this method need to be investigated in order to achieve similar results to the current state of the art methods that also employ linguistic analysis.

ACKNOWLEDGEMENTS

This research was supported by project No.264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1.

REFERENCES

Bollen, J., Rodriguez, M. A. and Van de Sompel, H. 2006. Journal Status. In: *Scientometrics 69 (3)*, pp. 669-687.

Brin, S.; Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In: *Computer Networks and ISDN Systems 30*: 107–117.

Chiru, C., Cojocaru, V. Trausan-Matu, S., Rebedea, T. and Mihaila, D. 2011. Repetition and Rhythmicity Based Assessment for Chat Conversation. In: *ISMIS 2011, LNCS 6804*, Springer, pp 513-523.

Hicklin, J., Moler, C., Webb, P., Boisvert, R., Miller, B., Pozo, R., Remington, K. Jama: a Java matrix package. (http://math.nist.gov/javanumerics/jama/ - accessed 16/08/2012).

Jiang, B. 2006. Ranking spaces for predicting human movement in an urban environment. In: *International Journal of Geographical Information Science 23 (7)*, pp. 823–837.

Muhlpfordt, M. and Wessner, M. 2005. Explicit referencing in chat supports collaborative learning. Paper presented at the *Proceedings of CSCL 2005*.

Navigli, R. and Lapata, M. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. In: *IEEE TPAMI, 32(4),* IEEE Press, pp. 678–692.

Page, L., Brin, S., Motwani, R., Winograd, T. 1998. *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report. Stanford InfoLab.

Rebedea, T., Dascălu, M., Trausan-Matu, Armitt, G., and Chiru, C. 2011. Automatic Assessment of Collaborative Chat Conversations with PolyCAFe. In: *Proceedings of ECTEL 2011, LNCS 6964*, Springer, pp. 299-312.

Schmidt, B. M. and Chingos, M. M. 2007. Ranking Doctoral Programs by Placement: A New Method. In: *PS: Political Science and Politics 40*, pp. 523–529.

Stahl, G. 2006. *Group cognition. Computer support for building collaborative knowledge*. Cambridge: MIT Press.

Stahl, G. 2009. *Studying Virtual Math Teams*. New York: Springer.

Tuulos, V. H. and Tirri, H., 2004. Combining topic models and social networks for chat data mining. In: *Proceedings of WI'04,* pp. 206–213.

Sundararajan, B., 2010. Emergence of the Most Knowledgeable Other (MKO): Social Network Analysis of Chat and Bulletin Board Conversations in a CSCL System. *Electronic Journal of E-Learning, 8(2),* pp. 191-207.

An Implementation-independent Evaluation Model for Server-based Signature Solutions

Thomas Zefferer and Bernd Zwattendorfer
Institute for Applied Information Processing and Communications, Graz University of Technology, Inffeldgasse 16a, 8010 Graz, Austria
{*thomas.zefferer, bernd.zwattendorfer*}*@iaik.tugraz.at*

Keywords: Electronic Signatures, Server-based Signature, Security Evaluation, Evaluation Model.

Abstract: During the past years, a general trend towards server-based signature solutions can be observed. Server-based signature solutions rely on a secure central server component that is able to securely store cryptographic keys and to create electronic signatures on behalf of users. Due to their various advantages compared to client-based solutions, it must be expected that server-based signature solutions will be increasingly deployed in security-critical fields of application in future. This raises the need for appropriate means to systematically evaluate the security of such solutions. Unfortunately, existing evaluation methods (e.g. Protection Profiles according to Common Criteria) are only partly applicable for evaluating server-based signature solutions. To overcome this issue, we propose a new implementation-independent evaluation model for server-based signature solutions. The proposed evaluation model is based on an abstract architectural model for server-based signature solutions and can hence be applied to arbitrary implementations. This way, we provide a powerful instrument to assess the security of future server-based signature solutions and pave the way for their adoption in security-critical fields of application.

1 INTRODUCTION

Electronic signatures have evolved to be an important instrument in online services. Depending on the given legal and organizational framework, the technical realization and implementation of electronic signatures is often no trivial task. For instance, in the European Union (EU) the EU Signature Directive (European Union, 1999) defines strict requirements for the creation of qualified electronic signatures, which are legally equivalent to handwritten signatures.

Considering European law, two basic approaches can be followed to implement solutions for the creation of qualified electronic signatures. *Client-based signature solutions* rely on a signature-creation token that is possessed, controlled, and maintained by the user, i.e. the signatory. Contrary, *server-based signature solutions* rely on a secure central server component that is able to store cryptographic keys and to create qualified electronic signatures on behalf of users.

During the past years, client-based signature solutions relying on smart cards have been the preferred implementation option especially for security-critical application scenarios. Unfortunately, client-based signature solutions usually lack an appropriate level of usability. This often leads to unsatisfactory user acceptance for applications based on qualified electronic signatures. Server-based signature solutions have the potential to overcome usability-related limitations of client-based approaches. An example is the server-based signature solution that has been introduced in Austria in 2010 (A-Trust, 2010). A conducted usability analysis has shown that this solution clearly outperforms existing client-based solutions in terms of user acceptance (Zefferer and Krnjic, 2012). Furthermore, this solution shows that server-based solutions are able to comply with existing legal and organizational requirements.

Due to the given advantages of server-based signatures, it can be expected that there will be an increasing number of server-based signature solutions in future. Hence, new server-based signature solutions will also be deployed in security-critical fields of application. This raises the need for appropriate means to systematically evaluate the security of server-based signature solutions in order to assess their suitability for security-critical fields of applications. Approved evaluation methods for signature solutions such as Common Criteria (Common Criteria, 2013) are already available. However, most of these methods are

only partly applicable for server-based signature solutions, as they usually do not properly consider their special characteristics.

To overcome this issue, we propose a new implementation-independent evaluation model for server-based signature solutions. The proposed evaluation model is based on an abstract architectural model for server-based signature solutions and can hence be applied to arbitrary implementations that comply with this architectural model. This way, we provide a powerful instrument to assess the security of future server-based signature solutions in order to assure their suitability and applicability in security-critical fields of application.

The remainder of this paper is structured as follows. In Section 2, basics of and related work on server-based signature solutions are discussed and their growing importance is emphasized. In Section 3, a generic architectural model for server-based signature solutions is introduced. Based on this model, the methodology that has been followed to develop the evaluation model proposed in this paper is presented in Section 4. Based on the presented methodology, an evaluation model for server-based signature solutions is developed and presented in Section 5. Finally, conclusions are drawn in Section 6.

2 SERVER-BASED SIGNATURES

The idea to rely on a trusted server component for the creation of electronic signatures on behalf of users is not new. Early concepts of server-based signature solutions have for instance been proposed in (Ding et al., 2002), (Bicakci and Baykal, 2003), and (Bicakci and Baykal, 2005). In general, server-based signature solutions have various advantages compared to client-based approaches. As computationally complex cryptographic operations are implemented by a server component, users do not need to acquire and maintain additional hardware or software in their client domain that is able to accomplish this task. Furthermore, server-based signature solutions can also be conceptually beneficial in terms of security, if security-critical cryptographic operations are carried out in a protected central environment such as a highly secure data-processing center.

Despite their conceptual advantages, server-based signature solutions have previously led a niche existence especially in security-critical application scenarios. The focus on client-based approaches in Europe was mainly caused by the EU Signature Directive (European Union, 1999), which defines the legal basis for electronic signatures in EU Member States. The EU Signature Directive defines that so-called *advanced electronic signatures* and *qualified electronic signatures*, which are legally equivalent to handwritten signatures, must be *'created using means that the signatory can maintain under his sole control'* (European Union, 1999). For several years, there has been common consent that this passage implies the use of user-owned signature-creation devices such as smart cards, even though this has never been explicitly postulated by this directive.

First discussions on alternative interpretations of the EU Signature Directive have been started in a *Public Statement on Server Based Signature Services* published by the *Forum of European Supervisory Authorities for Electronic Signatures (FESA)*[1]. This document explicitly supports the idea that server-based signature solutions comply with the requirements defined by the EU Signature Directive (Forum of European Supervisory Authorities for Electronic Signatures, 2005). In 2010, a server-based signature solution that supports the creation of qualified electronic signatures has been introduced and deployed in Austria (A-Trust, 2010). This solution has been based on a concept discussed in (Orthacker et al., 2010) and shows that server-based signature solutions are indeed feasible and applicable in security-critical application scenarios.

The growing importance of server-based signature solutions is also considered by legislative bodies and standardization committees. Documents published e.g. by the European Committee for Standardization (CEN) (European Committee for Standardization, 2013) and the European Commission (European Union, 2012) already take the possibility of server-based signature solution into account. The recent development of productive server-based signature solutions and the ongoing work on related standards and legal frameworks in Europe show that server-based signature solutions can be expected to further gain importance in the near future.

In order to provide appropriate means and methods to systematically evaluate the reliability and security of such solutions, we propose a new implementation-independent evaluation model for the security assessment of server-based signature solutions. To achieve implementation independence, the proposed evaluation model bases on a generic architectural model for server-based signature solutions, which is introduced in the following section.

[1] http://www.fesa.eu

3 ARCHITECTURAL MODEL

The evaluation model proposed in this paper focuses on solutions that require users to authorize signature-creation processes by means of a two-factor authentication. These solutions have gained importance during the past years, as they are able to provide the same level of security as client-based signature-creation methods relying on two factors. Server-based signature solutions relying on two-factor based user authentication typically make use of a secret password to cover the authentication factor *knowledge* and of one-time passwords (OTP) being sent to a device owned by the legitimate user[2]. By proving reception of the OTP, the user proves control over the device, which in turn implements the authentication factor *possession*.

An abstract model for a server-based signature solution that relies on the delivery of OTPs to implement two-factor based user authentication is shown in Figure 1. This model comprises two core components. The *System under Evaluation (SuE)* comprises all components of the server-based signature solution that are subject to the evaluation model proposed in this paper. The server-based signature solution itself makes use of and interacts with several external components. These components are subsumed under the abstract component *Environment*.

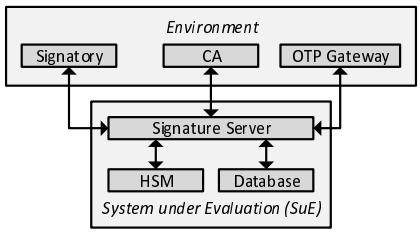


Figure 1: System model.

The SuE basically consists of three components. The *Signature Server* is the central component, which implements most functionality and interacts with external components, i.e. with the Environment, through well-defined interfaces. The Signature Server uses a *Database* to store required data such as user-related and user-identifying information. Additionally, the Signature Server uses a *Hardware Security Module (HSM)* to securely store security-critical data (e.g. signing keys) and to carry out cryptographic operations in a secure environment.

The Environment basically comprises three entities. The *Signatory* is the end user, who aims to create an electronic signature using the SuE. Users usually need to register once at the SuE prior to using its signature-creation functionality. During registration, a cryptographic key pair is generated by the SuE (more precisely by the HSM). While the private key remains at the SuE, the public key is sent to the *Certification Authority (CA)*, which issues a signing certificate for the Signatory. This certificate links the Signatory's identity to the generated key pair. After successful completion of the registration process, the Signatory can use the SuE to create electronic signatures.

To start a signature-creation process, the Signatory sends the *Data to be Signed (DTBS)* to the Signature Server. Before creating the signature, the Signature Server requests the Signatory to enter a secret password. Additionally, the Signature Server generates an OTP and transmits the OTP to the *OTP Gateway*. The OTP Gateway delivers the OTP to the Signatory. By entering the received OTP at the Signature Server, the Signatory completes the two factor-based user authentication process. After successful user authentication, the Signature Server loads required Signatory-related data from the Database and computes the electronic signature on the received DTBS in the HSM. Optionally[3], the Signature Server also displays the DTBS to the Signatory and requests an additional confirmation from the Signatory to sign these data. The computed signature is finally returned to the Signatory.

Based on this implementation-independent architectural model we propose an evaluation model for server-based signature solutions. The methodology followed to develop the proposed evaluation model is discussed in the next section.

4 METHODOLOGY

To rely on an approved method and to follow a systematic approach, we have based our evaluation model for server-based signature solutions on the concepts of Common Criteria (Common Criteria, 2013) and on the Protection Profile for Secure Signature Creation Devices (SSCD) Type 3 (CEN/ISSS, 2001). Development of security recommendations along the concepts of Common Criteria and Protection Profiles

[2]Other approaches to realize two-factor based user authentication are basically feasible. However, OTP-based approaches are currently wide-spread and already used in productive signature solutions such as (A-Trust, 2010). Hence, the evaluation model proposed in this paper focuses on OTP-based solutions only.

[3]In most cases, the applied legal framework determines whether DTBS must be displayed to the user during the signature-creation process or not.

is common practice. For instance, the Council of Europe has applied this methodology in the context of a risk analysis for e-voting solutions (European Council, 2004).

As Common Criteria in general and the Protection Profile for Secure Signature Creation Devices (SSCD) Type 3 in particular have not been developed with server-based signature solutions in mind, we have adapted underlying concepts where necessary to tailor our model to the special characteristics of server-based signature solutions. The resulting methodology that has been followed to develop the proposed evaluation model is shown in detail in Figure 2.

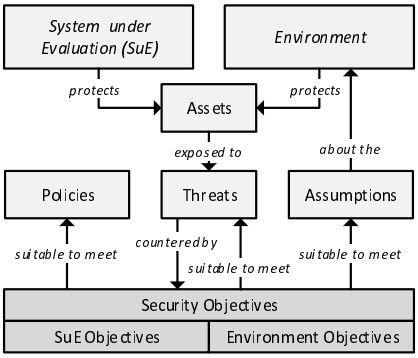


Figure 2: Methodology.

The top of the figure shows the two core components of server-based signature solutions according to the generic architecture introduced in Section 3: the *System under Evaluation (SuE)* and the *Environment*. Both components protect a set of *Assets*. Assets are values that need to be protected, e.g. secret cryptographic keys. Assets are exposed to *Threats*, which potentially compromise the assets' security. The set of potential threats is delimited by a set of *Assumptions* on the environment, in which the SuE is operated. Furthermore, potential threats are delimited by a set of *Policies*. From the made assumptions, defined policies, and identified threats, a set of *Security Objectives* is derived. The derived security objectives counter all identified threats and are also suitable to meet all made assumptions and defined policies.

According to this methodology, the proposed evaluation model for server-based signature solutions has been developed as follows. Based on the generic architecture shown in Figure 1, we have first defined a set of assets that need to be protected by server-based signature solutions. From these assets, potential threats have been identified that threaten to compromise the assets' security. Identification of potential threats has been based on a set of predefined assumptions on the environment and on a predefined set of policies. Assumptions, policies, and threats have then been used to derive security objectives for server-based signature solutions.

Concrete implementations of server-based signature solutions can be assessed with the proposed evaluation model by verifying if the given implementation meets all security objectives and is hence protected against all possible threats. Thus, the developed evaluation model is roughly comparable with a Protection Profile according to Common Criteria, whereas the identified security objectives of our model correspond to security requirements of typical Protection Profiles.

5 EVALUATION MODEL

In this section, we develop an evaluation model for server-based signature solutions following the methodology introduced in Section 4. To facilitate an understanding of analogies and relevant differences between the Protection Profile we rely on and the proposed evaluation model, we intentionally re-use notations that have been defined and introduced in (CEN/ISSS, 2001). In the following subsections, we develop the proposed evaluation model stepwise by following the methodology described above.

5.1 Assets

Assets define values that need to be protected. From the implementation-independent architectural model defined in Section 3, the following assets can be derived for server-based signature solutions:

- **A.1** *SCD:* SCD (signature-creation data) are private cryptographic keys, which are uniquely assigned to the Signatory (i.e. the user). These keys are used by the Signatory to create electronic signatures. SCD must always be kept confidential and protected from unauthorized access.
- **A.2** *SVD:* SVD (signature-verification data) are public keys, which are required to verify signatures created with the corresponding SCD. The integrity of SVD must be preserved.
- **A.3** *DTBS and DTBS display:* DTBS (data to be signed) represent the input to a signature-creation process, i.e. the data being signed by applying a cryptographic signature-creation function. The integrity of these data must be preserved before and during the signing process as well as during displaying these data to the Signatory in the course of the signature-creation process for confirmation purposes.
- **A.4** *VAD:* VAD (verification authentication data) need to be provided by the Signatory to authorize a signature-creation process. According to the used architectural model, VAD are represented by secret passwords. The authenticity and confidentiality of these data must be preserved.

A.5 *Signature-creation Function:* The cryptographic method used to create electronic signatures must comply with approved quality standards.

A.6 *Electronic Signature:* Created electronic signatures must be tamper-proof.

A.7 *OTP:* The confidentiality and integrity of one-time passwords (OTP) that are used by the Signatory together with her VAD must be preserved.

A.8 *User Data:* The integrity and confidentiality of Signatory-related data stored in the central database must be preserved.

A.9 *User ID:* The user ID unambiguously identifies the Signatory. This is necessary to select the correct SCD for each signature-creation process and to send the OTP to the correct user. The confidentiality and integrity of the user ID must be preserved.

A.10 *Database:* The database stores all required user-related data. The integrity of the database must be preserved and access to the database must be protected.

A.11 *Signature server:* The signature server provides functionality for accessing internal server components such as the database or the HSM. Physical and electronic access to the signature server must be protected and its integrity must be preserved.

A.12 *HSM:* The HSM (hardware security module) provides cryptographic functionality on a secure and reliable basis. Physical and electronic access to the HSM must be protected and its integrity must be preserved.

A.13 *HSM Master Key:* This is a secret key that is unique for each HSM, and which is used to protect the HSM's functionality and data. This key must always be kept confidential.

A.14 *Session ID:* The session ID is used to uniquely map a signature-creation operation to a specific Signatory and facilitates the parallel processing of requests from multiple Signatories. The integrity of the session ID must be preserved.

5.2 Assumptions

The following assumptions define properties and capabilities of the environment, in which the SuE is operated. The made assumptions delimit the set of potential threats for identified assets. All assumptions must be covered by appropriate security objectives.

AS.1 *CGA_Cert:* The Signature Server and the CA (certificate authority) are able to mutually authenticate each other.

AS.2 *CGA_Init:* The CA implements an appropriate registration and identification procedure and transmits relevant information to the Signature Server.

AS.3 *CGA_Secure:* The Signature Server has a secure channel to the CA, which is able to issue appropriate signing certificates according to given legal requirements and to verify that the user has access to her private keys.

AS.4 *OTP_Trusted:* The OTP Gateway is trusted and can be authenticated by the Signature Server.

AS.5 *OTP_Secure:* OTPs can be transmitted from the Signature Server to the OTP Gateway through a secure channel.

AS.6 *User_Trusted:* The Signatory uses a trustworthy end-user system to access signature-creation functionality provided by the Signature Server and to receive OTPs.

AS.7 *User_Secure:* The user communicates with the Signature Server through a secure channel.

AS.8 *Avail:* All external components are available during the registration and/or the signature-creation process.

5.3 Policies

Similar to assumptions on the environment, the definition of policies helps to delimit the set of potential threats. Again, all defined policies must be covered by appropriate security objectives. For the proposed evaluation model, the following policies have been defined:

P.1 *CA_QCert:* The CA is able to issue certificates for the Signatory's public key (SVD) according to the requirements defined by the relevant legal framework.

P.2 *QSign:* The Signatory uses an electronic-signature scheme that complies with given legal frameworks and requirements.

P.3 *Sig_System:* The Signatory's private key (SCD) is virtually unique.

P.4 *HSM_Secure:* The HSM is able to securely import and export data. The HSM does never export its secret master key. The HSM features secure cryptographic key generation and signing functions.

P.5 *HSM_Sig:* Electronic signatures are created inside the HSM only if all required data are available in the HSM and if all components of the SuE are in a defined state.

P.6 *System_Secure:* The SuE is deployed and operated in an appropriately secured environment and its components are protected from unauthorized access.

5.4 Threats

Based on the architectural model defined in Section 3, several threats can be derived that threaten to compromise the security of identified assets. Considering the above-defined assumptions and policies, the following threats can be derived:

T.1 *Hack_Phys:* An attacker accesses physical interfaces of the SuE to mount physical attacks such as side-channel analysis or fault attacks.

T.2 *SCD_Divulg:* An attacker gains access to confidential SCD stored by the SuE.

T.3 *SCD_Derive:* An attacker derives confidential SCD from public data such as SVD or electronic signatures created with the SCD.

Table 1: Assets targeted by threats.

	A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	A.9	A.10	A.11	A.12	A.13	A.14
T.1	X		X			X						X	X	
T.2	X				X	X								
T.3	X				X	X								
T.4			X			X								
T.5	X					X								
T.6		X												
T.7			X			X								
T.8					X									
T.9					X		X				X			
T.10							X				X			
T.11					X		X	X			X			
T.12									X					
T.13								X						
T.14								X	X	X			X	
T.15			X	X	X	X	X	X	X	X	X	X		X
T.16					X						X	X	X	
T.17	X				X						X	X	X	
T.18				X										
T.19			X		X	X		X						X
T.20			X		X	X		X						X
T.21			X		X	X	X			X	X	X		X

T.4 *Sig_Forgery:* An attacker forges signed data or the corresponding electronic signature.

T.5 *Sig_Repud:* An attacker compromises the non-repudiation of an electronic signature by modifying signature-relevant data.

T.6 *SVD_Forgery:* An attacker forges the Signatory's public key (SVD) during transmission to the CA in order to compromise the SVD's integrity.

T.7 *DTBS_Forgery:* An attacker modifies the DTBS that are displayed to the user. This way, the Signatory can be tricked into signing unintended data.

T.8 *SigF_Misuse:* An attacker misuses functionality provided by the system to sign arbitrary data on behalf of the legitimate Signatory without her knowledge.

T.9 *OTP_Forgery:* An attacker modifies or blocks delivered OTPs. This way, successful user authentication and hence signature-creation processes can be prevented by the attacker.

T.10 *OTP_Replay:* An attacker uses an intercepted OTP to authorize subsequent signature-creation processes.

T.11*OTP_Derive:* An attacker uses an intercepted OTP to derive useful information such as subsequent OTPs.

T.12 *UserID_Forgery:* An attacker modifies or deletes the unique ID of a Signatory to redirect the delivery of authentication data or to prevent successful signature-creation processes.

T.13 *Userdata_Forgery:* An attacker copies, modifies, deletes, or publishes user-related data to compromise their confidentiality.

T.14 *DB_Forgery:* An attacker steals or destroys the internal database of the SuE to compromise the system's integrity.

T.15 *Server_Forgery:* An attacker steals or destroys the Signature Server to compromise the system's integrity.

T.16 *HSM_Forgery:* An attacker steals or destroys the HSM of the SuE to compromise the system's integrity and to reveal confidential data.

T.17 *HSM_Compr:* An attacker compromises the master key of the HSM to gain access to assets stored inside the HSM.

T.18 *PIN_Compr:* An attacker gains access to the Signatory's VAD. This partly compromises the authentication scheme, which protects the Signatory's centrally stored private keys.

T.19 *Sess_Forgery:* An attacker guesses, modifies, or creates the ID of a session between the Signatory and the SuE. This enables an attacker to terminate an established session or to impersonate the Signatory.

T.20 *Sess_Hijack:* An attacker hijacks an established session between the Signatory and the SuE to gain the same privileges as the legitimate Signatory.

T.21 *System_Malfunc:* A malfunction of one or more components of the system leads to errors during the signature-creation process.

Each identified threat potentially applies to one or multiple assets. The concrete mapping between assets defined in Section 5.1 and threats defined in this section is provided in Table 1.

5.5 Security Objectives

A secure and reliable server-based signature solution must be able to counter all potential threats and meet all made assumptions and defined policies. From the identified threats, made assumptions, and defined policies listed above, we can therefore derive the following set of security objectives:

O.1 *Lifecycle_Security:* The SuE must be able to detect errors during initialization, personalization, and operation.

O.2 *SCD_Secrecy:* The confidentiality of the Signatory's private key (SCD) must be guaranteed.

O.3 *SCD_SVD_Corresp:* The SuE assures the correct relation between the Signatory's private and public key. Upon request, the SuE must be able to verify the relation between private and public key.

O.4 *SVD_Auth:* The SuE provides means that enable the CA to verify the authenticity of the Signatory's public key (SVD).

O.5 *Tamper_ID:* The SuE is able to detect physical manipulations of its components.

O.6 *Tamper_Resistance:* The SuE is resistant against manipulations of its components.

O.7 *Tamper_Response:* The SuE performs appropriate actions to protect components or data when tampering of the SuE has been detected.

O.8 *Init:* The SuE provides appropriate means to assure that private and public keys can be created by authorized persons only.

O.9 *SCD_Unique:* The SuE guarantees an appropriate level of quality for cryptographic keys. Private keys (SCD) must be virtually unique and must not be derivable from corresponding public keys (SVD).

Table 2: Threats, assumptions, and policies covered by security objectives.

Threats/Assumptions/Policies vs. Objectives	O.1 Lifecycle_Security	O.2 SCD_Secrecy	O.3 SCD_SVD_Corresp	O.4 SVD_Auth	O.5 Tamper_ID	O.6 Tamper_Resistance	O.7 Tamper_Response	O.8 Init	O.9 SCD_Unique	O.10 DTBS_Integrity	O.11 Sig_SigF	O.12 Sig_Secure	O.13 HSM_Secure	O.14 HSM_Trusted	O.15 HSM_Feature	O.16 HSM_Sig	O.17 Server_Trusted	O.18 System_Secure	O.19 VAD_Secure	O.20 DB_Access	O.21 DB_Secure	O.22 DB_Encrypt	O.23 DB_Bound	O.24 System_Avail	O.25 CA_QCert	O.26 SVD_Auth_CA	O.27 CA_Auth	O.28 VAD_Auth	O.29 VAD_Secure	O.30 User_Trusted	O.31 User_Secure	O.32 Avail
T.1 Hack_Phys		X			X	X	X																									
T.2 SCD_Divulg		X												X		X				X	X	X	X								
T.3 SCD_Derive									X			X		X																		
T.4 Sig_Forgery	X	X	X	X	X	X	X					X		X		X									X	X						
T.5 Sig_Repud	X	X	X	X	X	X	X		X	X	X	X		X		X				X	X	X	X		X	X						
T.6 SVD_Forgery				X																						X						
T.7 DTBS_Forgery										X			X	X			X													X	X	
T.8 SigF_Misuse										X	X		X	X		X	X			X	X	X										
T.9 OTP_Forgery																			X									X	X	X	X	
T.10 OTP_Replay																	X		X									X	X	X	X	
T.11 OTP_Derive																	X		X													
T.12 UserID_Forgery																	X			X	X	X										
T.13 Userdata_Forgery																	X			X	X	X										
T.14 DB_Forgery																		X					X	X								
T.15 Server_Forgery																	X	X						X								
T.16 HSM_Forgery																		X					X	X								
T.17 HSM_Compr														X																		
T.18 PIN_Compr													X								X										X	X
T.19 Sess_Forgery																	X							X							X	X
T.20 Sess_Hijack																															X	X
T.21 System_Malfunc																	X							X								
AS.1 CGA_Cert																											X					
AS.2 CGA_Init																									X							
AS.3 CGA_Secure																											X					
AS.4 OTP_Trusted																												X				
AS.5 OTP_Secure																													X			
AS.6 User_Trusted																															X	
AS.7 User_Secure																																X
AS.8 Avail																																X
P.1 CA_QCert			X																						X							
P.2 QSign											X	X													X							
P.3 Sig_System								X	X		X																					
P.4 HSM_Secure													X	X	X						X	X										
P.5 HSM_Sig														X		X																
P.6 System_Secure																		X														

O.10 *DTBS_Integrity:* The SuE must assure that DTBS are not modified while transmitted to the HSM. Furthermore, DTBS must not be modified when being displayed to the Signatory. The SuE must assure that exactly the same DTBS are displayed and signed by the HSM.

O.11 *Sig_SigF:* The SuE provides signature-creation functionality to legitimate users only and protects a Signatory's private key from access and use by others.

O.12 *Sig_Secure:* The SuE creates electronic signatures, which cannot be forged without knowledge of the private key (SCD). The private key cannot be derived from created signatures.

O.13 *HSM_Secure:* The HSM provides appropriate means to securely transmit security-critical data such as SCD or DTBS.

O.14 *HSM_Trusted:* The trustworthiness of the HSM and its cryptographic operations must be guaranteed.

O.15 *HSM_Feature:* The HSM must be designed and implemented such that an error-free and correct operation is assured.

O.16 *HSM_Sig:* The HSM is able to carry out all required cryptographic operations in a protected environment.

O.17 *Server_Trusted:* The signature-server component must be implemented according to its specification and must process signature-creation requests as expected.

O.18 *System_Secure:* It must be assured that the SuE and all of its components are protected against physical and electronic intrusion and that signature-creation processes are carried out correctly.

O.19 *VAD_Secure:* Authentication data (VAD) that are used to authenticate Signatories prior to signature-creation processes must be protected against misuse by unauthorized persons.

O.20 *DB_Access:* Access to the SuE's database must be restricted to authorized persons and components.

O.21 *DB_Secure:* Data transmissions to the database from other components must rely on secure communication channels to prevent eavesdropping and modification of transmitted data.

O.22 *DB_Encrypt:* The SuE must assure that user-related data stored in the database can be accessed by the correct user and can be used for signature-creation processes of this user only.

O.23 *DB_Bound:* Private keys (SCD) must be bound to the HSM. Furthermore, the use of private keys must be limited to the signing of DTBS.

O.24 *System_Avail:* Technical and organizational means must be in place to assure the availability of the SuE

and to guarantee that no data are lost in case of system errors or crashes.

O.25 *CA_QCert:* The CA creates certificates according to the given legal requirements.

O.26 *SVD_Auth_CA:* The CA verifies the integrity of the obtained public key (SVD), checks the origin of the public key, and verifies the relation between public key and private key.

O.27 *CA_Auth:* The CA must authenticate at the SuE. Furthermore, the SuE and the CA must communicate over an authenticated and secure channel.

O.28 *VAD_Auth:* The OTP Gateway must be able to prove its trustworthiness.

O.29 *VAD_Secure:* The SuE and the OTP Gateway must communicate over an authenticated and secure channel.

O.30 *User_Trusted:* The Signatory must assure that the used end-user system is trustworthy and free from malware and other sources of interference.

O.31 *User_Secure:* The Signatory must assure that her communication with the SuE relies on a secure and trustworthy channel.

O.32 *Avail:* Technical and organizational means must be in place to assure the availability of external components.

By meeting the derived security objectives, which finally represents the developed evaluation model, server-based signature solutions are able to counter all identified threats and meet all made assumptions and defined policies. Each security objective covers one or multiple threats, assumptions, or policies. The concrete mapping between security objectives, threats, assumptions, and policies is provided in Table 2.

6 CONCLUSIONS

In this paper we have proposed an evaluation model for the systematic assessment of arbitrary server-based signature solutions. The proposed model basically defines a set of implementation-independent security objectives. Security objectives have been derived following an elaborate methodology aligned upon the approved concept of Common Criteria. The security of a concrete server-based signature solution can be assessed with the help of the proposed evaluation model by determining its capability to meet the set of defined security objectives.

Application of the proposed evaluation model to existing server-based signature solutions is regarded as future work and will kill two birds with one stone. First, the soundness of the proposed evaluation model will be evaluated. Second, existing server-based signature solutions will be systematically assessed in order to identify potential security vulnerabilities. By providing a universal evaluation model for arbitrary implementations, the proposed evaluation model helps to assess and assure the security of future server-based signature solutions. This facilitates a future adoption of server-based signature solutions also in security-critical fields of application and paves the way for secure *and* usable online services based on electronic signatures.

REFERENCES

A-Trust (2010). Activate mobile phone signature. http://www.buergerkarte.at/en/activate-mobile.html.

Bicakci, K. and Baykal, N. (2003). Saots: A new efficient server assisted signature scheme for pervasive computing. In Hutter, D., Mller, G., Stephan, W., and Ullmann, M., editors, *SPC*, volume 2802 of *Lecture Notes in Computer Science*, pages 187–200. Springer.

Bicakci, K. and Baykal, N. (2005). Improved server assisted signatures. *Computer Networks*, 47(3):351–366.

CEN/ISSS (2001). Protection profile - secure signature creation device type 3. http://wwww.commoncriteriaportal.org/files/pp files/pp0006b.pdf.

Common Criteria (2013). Common criteria. http://www.commoncriteriaportal.org/.

Ding, X., Mazzocchi, D., and Tsudik, G. (2002). Experimenting with server-aided signatures. In *NDSS*. The Internet Society.

European Committee for Standardization (2013). Security requirements for trustworthy systems supporting server signing. https://shop.austrian-standards.at/Preview.action?preview=&dokkey=478405.

European Council (2004). Multidisciplinary ad hoc group of specialists on legal, operational and technical standards for e-enabled voting (ip1-s-ee) b. explanat.

European Union (1999). Directive 1999/93/ec of the european parliament and of the council of 13 december 1999 on a community framework for electronic signatures.

European Union (2012). Proposal for a regulation of the european parliament and of the council on electronic identification and trust services for electronic transactions in the internal market.

Forum of European Supervisory Authorities for Electronic Signatures (2005). Public statement on server based signature services. http://www.fesa.eu/public-documents/PublicStatement-ServerBasedSignatureServices-20051027.pdf.

Orthacker, C., Centner, M., and Kittl, C. (2010). Qualified mobile server signature. In *Proceedings of the 25th TC 11 International Information Security Conference*.

Zefferer, T. and Krnjic, V. (2012). Usability evaluation of electronic signature based e-government solutions. In *Proceedings of the IADIS International Conference WWW/INTERNET 2012*, pages 227 – 234.

Business Process Improvement Methodology Based on Business Process Modelling

Case Study from Healthcare Sector

Faisal Aburub
Faculty of Administrative and Financial Sciences, MIS Department, University of Petra, Amman, Jordan
faburub@uop.edu.jo

Keywords: Business Process, Business Process Improvement, Process Modelling.

Abstract: Business process can be considered as one of the key pillars of successful e-business. One of the most important objectives of e-business is to support an organization by automating its business processes using information systems. It is important to make sure that the anticipated information systems is economically feasible and support business process of such organization effectively and sufficiently otherwise, the information systems will fail. One reason of information systems and e-business fail is because there are deficiencies in business processes of an organization. Therefore, a new methodology has been developed for business process improvement based on business process modelling particularly, Role Activity Diagramming (RAD). The methodology has been validated and tested using cancer registration process in Jordan as a case study. The key features of the methodology are that it is grounded in process modelling and brings together functional, non-functional and structural process analyses. Further research work is being undertaken to replicate this study on other application domains and thus generalize the adopted approach.

1 INTRODUCTION

According to Chaffey (2011), e-business is a wider concept that embraces all aspects of the use of information technology in business. It includes not only buying and selling, but also servicing customers and collaborating with business partners, and often also involves integration across business processes and communication within the organization. Therefore, business process can be considered as one of the key pillars of successful e-business. One of the most objectives of e-business is to support an organization by automating its business processes using information systems. Unfortunately, developing information systems is sometimes difficult and expensive process as it needs to go through system development life cycle. So, it is important to make sure that the anticipated information systems is economically feasible and support business process of such organization effectively and sufficiently otherwise, the information systems will fail.

One reason of information systems and e-business fails is because of there are deficiencies in business processes of an organization. Therefore, business process precedes information systems. This paper aims to develop a methodology for business process improvement based on business process modelling. The methodology will be validated and tested using cancer registration process in Jordan as a case study. Process modelling can be used to provide a comprehensive understanding of business activities and functions and thence a base for detailed process analysis. Business process improvement, on the other hand, refers to a family of approaches which aim to help an organization to adjust its processes to fit a dynamic or complex business environment, particularly so as to take advantage of rapid advance in information technology. In particular, business process improvement methods exist in order to identify deficiencies and problems that occur frequently and prevent organizations from achieving their goals. Although many methodologies have been developed to improve business processes, most of them do not make significant use of process modelling to guide the evaluation and improvement of business processes. For example, (Sun and Liu, 2010), (Pyona et al., 2011), (Milan et al., 2013), (Reza and Srinivasan, 2013), (Botha et al., 2012) and (Damija et al., 2008) have developed methodologies to improve business processes without significant role for business process modelling. (Jose, 2012)

suggested a multi-model methodology that implements a smooth and continuous process improvement, depending on the organization's business goals and allowing users to establish their improvement implementation pace. (Borger, 2012) proposed six criteria which can help to recognize practical and reliable tool-supported business process description and modeling systems. These criteria had been proposed based on the investigation three approaches describing models of business processes: the OMG standard BPMN in its recent version 2.0, the workflow patterns of the Workflow Pattern Initiative and their reference implementation YAWL. (Mast et al., 2011) proposed a methodology for Process Improvement in Healthcare. This methodology based on the connexion between standard models from operation management and axiological model which breaks down general performance goals into process metrics. Business process re-engineering was introduced originally by (Hammer, 1990) and (Davenport and Short, 1990). (Nicola et al., 2012) proposed the synergic use of Business Process Abstract Language, a BP modeling framework to represent BP schemas, and Object Process Actor modeling Language, an ontological framework to capture the semantics of a business scenario. (Michele, 2010) identified new reengineering methodology called HY-CHANGE which is conceived as a logical and technical support to the decision maker. It results in a number of recursive phases, where the rational and synchronous use of several tools and techniques, borrowed from business process re-engineering (BPR) and continuous quality improvement (CQI) principles, is structured into a unique approach.

The current research uses process modelling techniques in a systematic and generalisable manner to gain deeper understanding of processes in a particular complex case. By analysis and further probing of the process models, it then seeks to develop a practical methodology for business improvement which will be applicable not only in the case in question but also more broadly. The methodology combines a number of approaches. The approaches include linking business objectives to finer grain activities (or tasks) in Role Activity Diagram (RAD) models (Ould, 2005) and systematic questioning of elements of the model to test validity and performance in terms of the overall goal of business process improvement. Furthermore, the concept of Non-Functional Requirements (NFRs) is applied to the analysis of the RAD business process models. This is an adaptation of a method developed in software engineering for linking NFRs to conceptual models. Translated into the domain of business process modelling, NFRs are equated with the general or overall quality attributes of a business process. These, though essential aspects of any effective process, are not well captured in a functionally oriented process model.

Section 2 provides a brief overview on business process improvement methodologies. Section 3 presents business process modeling. This is followed by proposed business process improvement methodology in section 4. Improving the JCR process is presented in section 5. Conclusion and future work are presented in section 6.

2 PROCESS MODELLING

This research uses the process modeling language RAD (Role Activity Diagramming) as one of its main techniques. Role Activity Diagrams (RADs) are diagrammatic notations to represent and model coordinated behaviour and interactions within a process. The advantage of the RAD modelling language for the present research lies in the relative ease of understanding of the diagrams produced, and the clear identification in the RAD models of the main roles (responsibilities), activities, decision points, and relationships in a business process. Also, the role, as represented in RAD is similar to the idea of a *module*, since it allows a grouping of primitive activities to be assigned to a particular person or agent. RADs express the modularity of work through roles and synchronisation between roles. The ability to express this modularity and responsibility make RADs a useful vehicle for performing assessments of a business process, since the structure of the work is clearly laid out.

3 CANCER REGISTRATION IN JORDAN

Cancer registration in Jordan is managed by the Jordan Cancer Registry (JCR) established in 1996 as a population-based registry as a unit of the Ministry of Health. Malignant and some benign cases have been recorded by JCR for people living in Jordan since January 1996.

JCR has been used in this research as a case study in which to investigate practically the use of business process modelling as an approach to improve understanding of processes and subsequently to achieve business process

improvement through redesign.

4 THE METHODOLOGY FOR IMPROVING BUSINESS PROCESSES

According to (Bendell, 2005), "business process improvement is fundamental to business development, quality improvement and the management of change". Business process improvement method is a structured approach to the analysis and continual improvement of the fundamental activities of an organisation by simplifying and streamlining its business processes (Lee and Chuah, 2001). Similarly, the new methodology now proposed has been developed as a means of analysing and improving complex processes through the use of different techniques for modelling and analysing current processes.

The methodology consists of seven phases namely, understand the current process, model the current process, validate the current process models, analyse and evaluate the current process, redesign the current process, evaluate the redesigned process, implement the redesigned process. The methodology will be explored in details in the next sections, and applied to the cancer registration process.

4.1 Phase I: Understand the Current Process

Understanding the current process is an essential step in business process improvement. Good understanding of business processes will feed and inform the next phase, particularly the process modelling phase.

4.2 Phase II: Model the Current Process

This phase can be considered as an extension of the understanding phase (first phase), since modelling is an important tool to assist in understanding and analysing business processes and their associated activities in an organisation. The aim of this phase is to model the current business processes using RAD.

4.3 Phase III: Validate the Current Process Models

The validation phase checks the RAD models of the current process to ensure that they represent an adequate description of the process. The models should accurately and comprehensively represent the roles, activities, interactions and workflow in the current process. Because RAD process models give a reasonably transparent and comprehensible picture of business processes, it is feasible to validate them through direct discussion with staff members in the organisation. This discussion can be with individuals and/or groups.

4.4 Phase IV: Analyse and Evaluate the Current Process

The analysis takes two forms: structural analysis of the models, and model-led extended investigation process.

4.4.1 Structural Analysis of the Current Process Models

Analysis of the RAD models themselves can give an initial indication where there might be unnecessary complexity, unvalidated data, crossovers between functions, and/or opportunities for IT support in the modelled business processes.

(1) Discovery of Unnecessary Complexity
RAD models can be used to discover how well the current process has achieved its goal. Interactions in the RAD models show how roles in a process coordinate to achieve the overall goal. By focusing on particular roles, we can find out how many activities a role comprises, and how many interactions it drives, by examining the current process model and its submodels. It may be possible to improve a certain process by redefining or reorganising roles, activities, or interactions if the present arrangements appear to cause unnecessary complexity, or reduce the quality of the process objectives.

(2) Discovery of Unvalidated Data
RADs can be used to ascertain whether collected data is being validated by examining the interactions that represent sending and receiving information between roles.

(3) Discovery of Cross-functional Boundaries
One essential objective of a business process is to simplify or reduce cross-functional flows within the overall process. The concepts of coupling and cohesion from the field of software systems design can be adapted in business process modelling for checking interfunctional connections. In relation to business processes, we may argue that a role should be highly cohesive, which implies that its activities

are closely integrated and serve a single purpose. (Ould, 1995) argues that roles should ideally be loosely coupled, so that the number of interactions between them is reduced to a minimum. If roles lack cohesion and/or are highly coupled, then this may cause excessive cross-functional connections, which can lead, according (Hindle, 1997), to excessive delays, inconsistencies, duplication of effort, loss of information, and customer dissatisfaction.

(4) Identification of Automated and Human-based Activities

It might be possible to use RAD models to identify the potential areas for automation, or IT support. It can be useful to indicate on a model, e.g. by using different colours, which activities are or could be automated. This could help measure the operational cost in a process by indicating the relative needs for human- and computer-based activities. More generally, the models could be annotated with additional information on other operational aspects, including resources, timings, volumes and methods, which are not covered in the basic notation.

Though structural analysis of the models does not cover all aspects of business processes, it gives an initial indication of problems and opportunities, so as to make a good start for a detailed evaluation.

4.4.2 Model-led extended Investigation Process

This second method of analysis involves querying each element (each activity and each interaction) of the process models, to gauge their level of efficiency. This involves a second round of questionnaires and interviews with organisational actors.

(1) Aspects for Investigation

Systematic, element-by-element evaluation of the model with the participants in the study can reveal deficiencies that prevent a business process from achieving its objectives.

Analysis of the current process can be achieved through developing targeted questions related to activities, interactions, and roles included in the models. These questions should be answered by informants, who can be customers, staff members, or both. The answers, which may include suggestions from role holders, can then help to discover process deficiencies. The questions should be developed in a systematic way according to the roles and workflows in the RAD models.

The element-by-element investigation just described will not readily capture high-level or general attributes of the business process as a whole. The RAD models essentially provide a functional view of the business process. By analogy to non-functional requirements in software systems, which represent critical quality attributes of such systems (Kotonya and Summerville, 1998), (Cysneiros et al, 2001) and (Chung and Nixon, 1995), we suggest it will be useful to identify and model such general desirable qualities or properties of business processes. An approach has been developed to model business process non-functional requirements (NFRs) and links them to RAD models, as a way of representing and assessing the general qualities of a business process. Our way of proceeding is to model the NFRs of a process and map the NFR model to the functional process model. Effectively, the NFR model is applied to the functional model by using it to identify activities, interactions, and roles in the functional model which could usefully be modified to deliver or serve the NFRs better. By systematic mapping of the NFR model to the functional process model, activities and interactions are modified so as to embed the NFRs within them.

(2) Qualitative and Quantitative Evaluation of the Current Process

Following development of the questions for the extended investigation, as described above, the next step is to collect and analyse the additional data from the informants.

4.5 Phase V: Redesign the Current Process

The purpose of this phase is to improve the current business process using the results of the previous phases, with the ultimate objective of achieving the business goals identified at the outset.

As indicated previously, general high-level qualities or NFRs are essential elements in a successful process, which could satisfy them through linking them to functional features. Therefore, an important prerequisite for process improvement is the proper representation and identification of NFRs.

4.5.1 Modelling NFRS of the Business Process

We have adapted Cysneiros and Chung's approach to NFR analysis, which they use in software engineering to resolve conflicts between goals in software systems (Chung, 1993), (Chung and Nixon, 1995) and (Cysneiros et al., 2001). To model NFRs for business processes, we propose the following six-step method: elicit NFRs and decompose them into sub-goals, define relationships between goals and sub-goals, identify actors who will achieve

NFRs (goals), operationalise goals, analyse positive and negative interactions between goals, and select which NFRs to address to achieve process improvement.

4.5.2 Linking Non-functional Requirements to Functional View

Linking the non-functional view represented in the NFR graphs to the functional view represented in the RAD models can be performed through the following steps: find all the NFR graphs that refer to a particular process in the RAD model, find all the NFR graphs that refer to a particular role in the RAD model, relate the activities and interactions included in this role to the roots of NFRs which correspond to the same role, if this role in the RAD model does not cover the nodes in the NFR graphs, then this RAD model must be updated by adding, modifying, or deleting activities or interactions that belong to that role, and if a role identified (as an actor) in the NFR graphs does not exist in the current RAD model, then a new role must be instantiated to satisfy that NFR.

4.6 Phase VI: Evaluate the Redesigned Process

The present phase evaluates the redesigned process models, prior to implementation, and considers the impact of the new models on the current process. (Kueng and Kawalek, 1997) state that process modelling can assist in the detection of problems in business processes. Therefore, it is possible to make an initial assessment of how good the new design is before starting to implement the new process. For example, one way to evaluate the newly designed processes is by checking level of automated activities.

4.7 Phase VII: Implement the Redesigned Process

RAD models of the redesigned process will provide a detailed blueprint for guiding the implementation of the new process, which will also aid communication between the improvement team and those involved in operating the new process. The RAD models will clearly illustrate the roles of the staff members, the changes in those roles, and the interactions between those roles. Suitably marked up, they can also indicate the points at which there will be some technological innovation. After implementing the redesigned process, the new process will need to be evaluated in order to measure whether the expected improvements have been achieved.

5 IMPROVING THE JCR PROCESS USING NEW METHODOLOGY

5.1 Understanding and Modelling the JCR Process

The main objective of the JCR process is to improve the administration of cancer treatment and the collection of information about cancer cases. Information about the current JCR process has been elicited using semi- and un-structured interviews conducted with staff members, and non-participant observation, performed in the Jordan Cancer Registry (JCR). The current JCR process has been modelled using RAD as shown in Figure 1.

5.2 Validation of the Rad Models of the JCR Process

Individual and group meetings were conducted to validate the first version of the model were carried out in JCR. Comments received during these meetings which prompted changes to the first RAD models, which led to the current version of the RAD model. These meetings were also used to identify and validate non-functional requirements (quality attributes) such as information quality.

5.3 Analyse and Evaluate the JCR Process

Analysis of RAD model and responses of the JCR staff has identified deficiencies and shortcomings in the cancer registration process as follows: JCR information is unvalidated, the computer system is generally helpful, but there are some problems and the system needs improvement, the available information is enough to do current analysis and reports but not any additional reports, the users are satisfied with the system performance, there is high volume of complex information and a long time is needed to collect the required information, and the current work flow needs some improvements, (Rub et al., 2008).

5.4 Redesign the JCR Process

The redesigned cancer registration process is shown as a RAD model in Figure 2. In this model, data collection in the JCR will be improved by electronic submission of data to JCR. The redesigned process will in addition be able to perform survival analysis, treatment evaluation, cancer incident reporting, data filtering according to international rules of cancer registration, cancer screening analysis, and follow up of patients. Furthermore, most of required NFR (qualities) are satisfied in the redesigned model of the JCR process. For example, information quality is satisfied in the redesigned model of the JCR process by the addition of appropriate interactions and activities such as 'accuracy of records and details', 'comparison JCR's data against different laboratories' and hospitals' data' (Aburub et al, 2007). The redesigned model and suggested improvements for the cancer registration process have been validated by managers at JCR, who indicated that the suggested improvements would be useful for JCR and could help the cancer registration process achieve its goals.

6 CONCLUSIONS AND FUTURE WORK

One of the important aims of e-business is to enhance business processes in the organisations. One method to achieve that, information technology should be in line with the business process. This research investigation has used process modelling and case study to deepen understanding of complex organizational process in order to make suggestions for business process improvements such as suitable investment in IT. From that, a general methodology for understanding and improving business processes has been developed. Furthermore, this research has extended process modelling to include non-functional or quality attributes of business processes. Seven stages have been proposed for the overall methodology. These stages have been applied and studied in the cancer registration case.

Furthermore, Role Activity Diagramming (RAD) was used as the main modelling tool. The resultant models were then subjected to a structural analysis and used as the basis for an extended investigation driven from the models. Other modelling methods

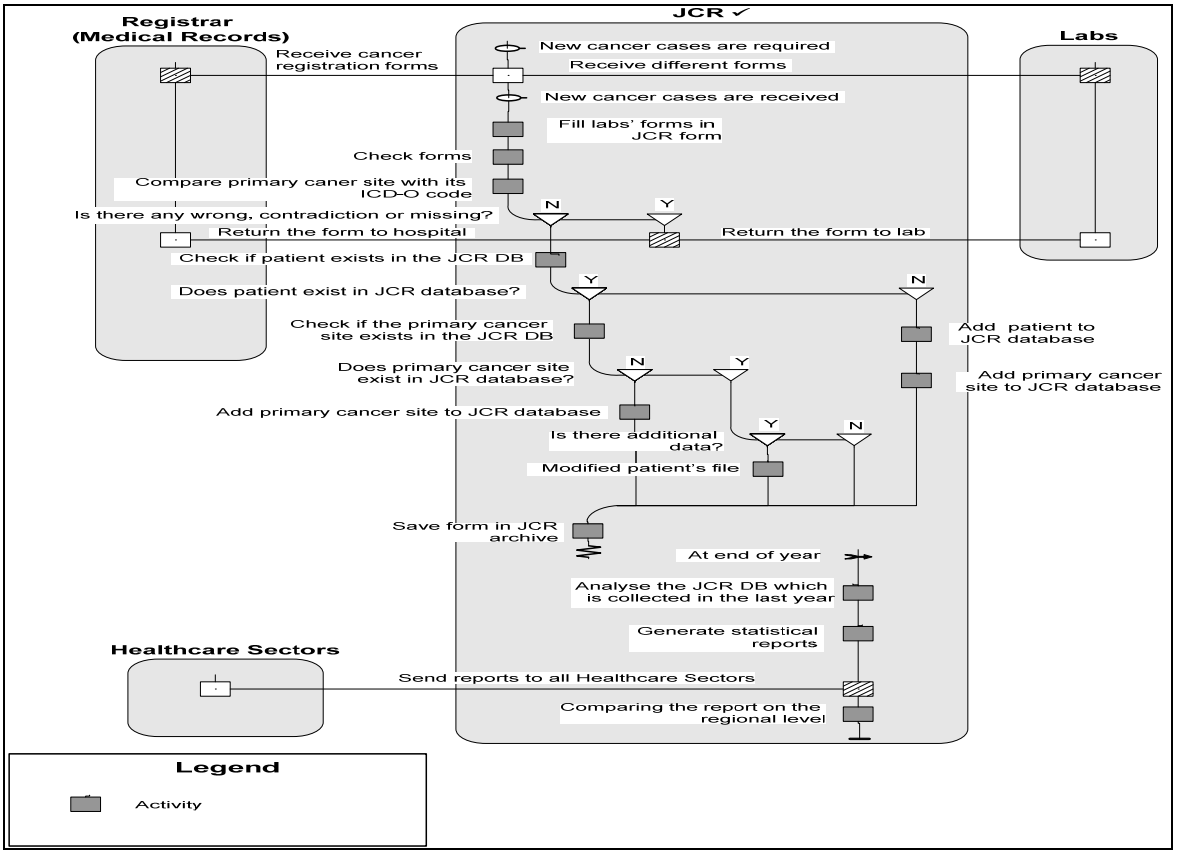


Figure 1: RAD model of current cancer registration process.

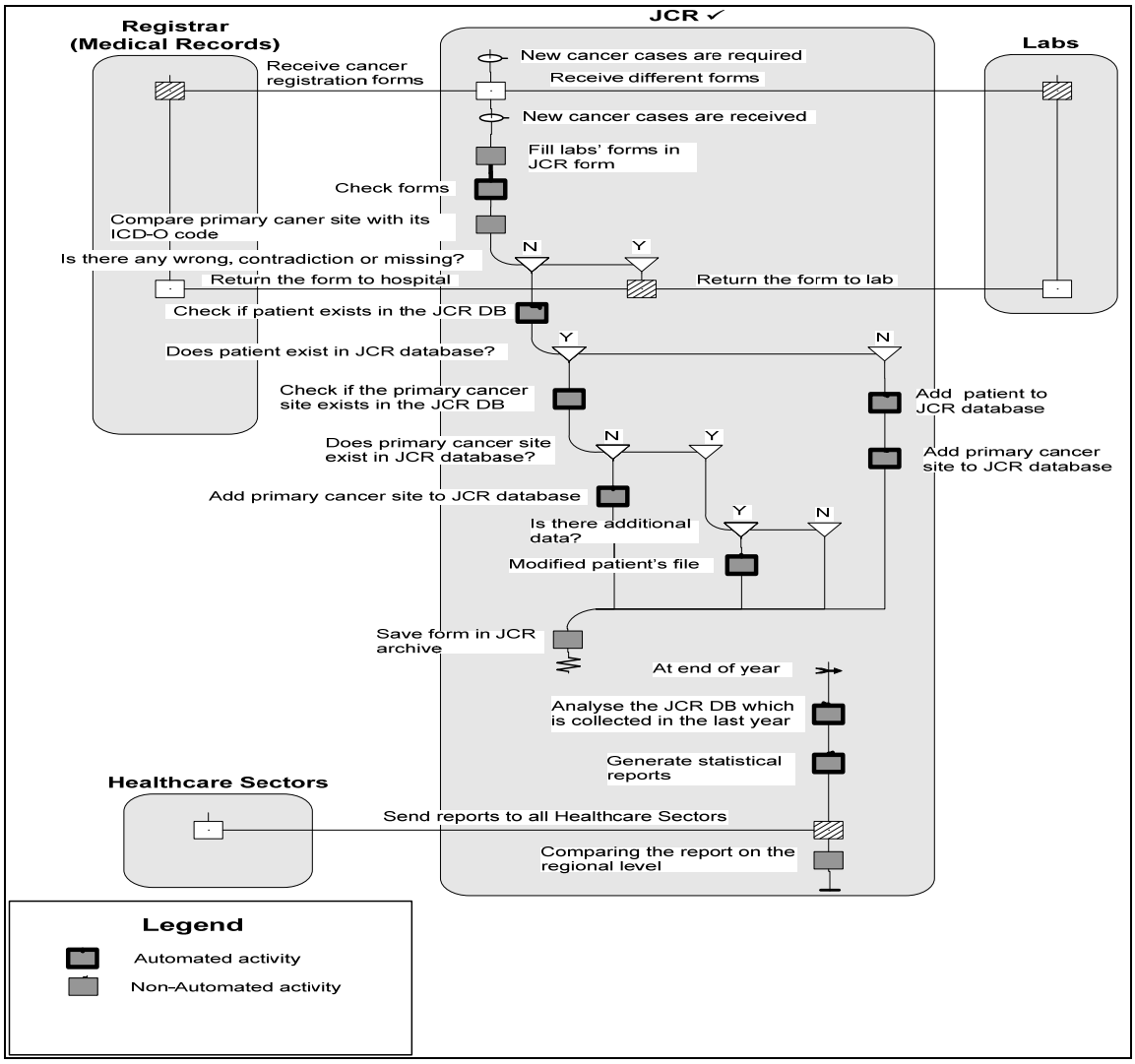


Figure 2: Improved RAD model of cancer registration process.

were used in addition, NFR graphing in particular. Process modelling was indeed thus used in this study to understand and analyse the business process of an organization.

Finally, the methodology that has been developed in this research has been investigated and critically evaluated using the cancer registration process in Jordan. Although it is reasonable to claim that the application of methodology using the cancer registration process has been successful, further application and test, especially in different domains would further confirm its general validity and usefulness.

REFERENCES

Chaffey, D., 2011. *E-Business and E-Commerce Management Startegy, Implemetation and Practice*. Pearson, 5th edition.

Sun Y., Liu X., 2010.Business-oriented software process improvement based on CMMI using QFD. *Information and Software Technology*. 52(1), pp 79-9.

Pyona C., Woob J., Parkc, S., 2011. Service improvement by business process management using customer complaints in financial service industry. *Expert Systems with Applications*. 38(4), pp 3267-3279.

Damija, N., Damijb, T., Gradc, J., Jelencd, F., 2008. A methodology for business process improvement and IS development. *Information and Software Technology*. 50(11), pp 1127-1141.

Milan R., Milan P., Nemanja B., Neskovic N., Branislav N., 2013. Reengineering of Supply Chain Process in Production Systems -- A Case Study. *Engineering Economics*.24 (1), pp71-80.

Reza P., Srinivasan K., 2013. The synergy of continuous process improvement. *Industrial Engineer: IE.* 45(6), pp 41-45.

Botha J., Kruger S., de Vries M., 2012. Enhancing Customer Experience Through Business Process Improvement: An Application Of The Enhanced Customer Experience Framework (Ecef). *South African Journal of Industrial Engineering*. 23 (1), pp 39-56.

Jose C., Gonzalo C., Gerzon G., Jezreel M., Mirna M., Tomás San F., 2012. Methodology For Process Improvement Through Basic Components And Focusing On The Resistance To Change. *Journal of Software: Evolution & Process*. 24(5), pp 511-523.

Borger E., 2012. Approaches to modeling business processes: a critical analysis of BPMN, workflow patterns and YAWL. Software System Model. 11, pp 305–318.

Mast J., Kemper B., Does R., Mandjes M., der Bijl Y., 2011. Process Improvement in Healthcare: Overall Resource Efficiency. *Quality and Reliability Engineering International*. 27, pp 1095 – 1106.

Hammer, M., 1990. Reengineering work: don't automate, obliterate. *Harvard Business Review,* 90 (4), pp 104 - 112.

Davenport, T., H., Short, J., E., 1990. The new industrial engineering: information technology and business process redesign. *Sloan Management Review*, 31(4), pp 11-27.

Nicola1 A.,Missikoff M., Smith F. 2012. Towards a method for business process and informal business rules compliance. *Journal Of Software: Evolution And Process.* 24, pp 341–360.

Michele D., 2010. HY-CHANGE: a hybrid methodology for continuous performance improvement of manufacturing processes. Journal of Production Research. 48 (15), pp 4397- 4422.

Bisson, B., Folk, V., Smith, M., 2000. Case study: how to do a business process improvement. *Journal for Quality and Participation,* 23(1), pp 58 – 63.

Bendell T., 2005. Structuring Business Process Improvement Methodologies. *Total Quality Management.* 16(8), pp. 969–978.

Lee, K. T., Chuah, K. B., 2001. A super methodology for business process improvement: an industrial case study in Hong Kong/China. *International Journal of Operations & Production Management*, 21(5), pp 687-706.

Ould M., 2005. *Business process management, a rigorous approach*. Swindon: British Computer Society.

Saven, S., 2003. Business process modelling: review and framework. *International Journal of Production Economics,* 90(2), pp 129-149.

Ould, M. A., 1995. *Business processes modelling and analysis for re-engineering and improvement*. Chichester: Wiley.

Hindle, J., 1997. Understanding business processes. *Health Manpower Management,* 23(5), pp 181-183.

Phalp, K., Shepperd, M., 2000. Qualitative analysis of static models of processes. *The Journal of Systems and Software, 52*, pp 105 – 112.

Kotonya, G., Sommerville, I., 1998. *Requirements engineering.* Chichester: Wiley.

Cysneiros, L., Leite, J., Neto, J., 2001. A framework for integrating non-functional requirements into conceptual models. *Requirements Engineering*, 6, pp 97 – 115.

Chung, L., Nixon, B., 1995. Dealing with non-functional requirements: three experimental studies of a process-oriented approach. In: *Proceedings of the 17th international conference on software engineering, Seatle, WA, April 1995*, pp 24 – 39.

Chung L., 1993. Representing and using non-functional requirements: A process oriented approach. Ph.D. Thesis, Dept. of Computer Science, Univ of Toronto.

Kueng, P., and Kawalek, P., 1997. Goal-based business process models: creation and evaluation. *Business Process Management Journal*, 3(1), pp 17-38.

Aburub, F., Odeh, M., Beeson, I., 2007. Modelling non-functional requirements of business processes. *Information and Software Technology*. 49 (11), 1162-1171.

Rub, F. A., Odeh, M., Beeson, I., Pheby, D., & Codling, B., 2008. Modelling healthcare processes using role activity diagramming. *International Journal of Modelling and Simulation*. 28(2), pp 147–155.

A Model to Specify Inter-company Cooperation

Meriem Kermani[1], Mahmoud Boufaida[1] and Emmanuel Paviot-Adet[2]
[1]LIRE laboratory, Constantine 2 University, Constantine, Algeria
[2]LIP6 laboratoty, Pierre & Marie Curie University, Paris, France
kermanimeriem@hotmail.fr, mboufaida@umc.edu.dz, emmanuel.paviotadet@lip6.fr

Keywords: Interoperability, Inter-company Cooperation, Learning Agents.

Abstract: The impact of the evolution of the global economic ecosystem, has forced and still force again organizations to either adopt new behaviour schemas and to deeply change their structure, also to be more open to their environment as well. These impacts require a new cooperation philosophy from organizations side. Several studies have focused on the problem of the intercompany cooperation, proposing approaches that provide interoperability mechanisms. But this remains an open research domain. In this work we will propose a mediation architecture between different companies. A mediator allows us to create an intercompany cooperative process, the purpose of this solution is to keep the company architecture and ask the mediator which is a software-based agent to play an intermediary role between companies, and involve it in making the transformation between companies. We define a dynamic and cooperative inter-companies model that combines the agent technology and the decision trees paradigm. This last facilitates making decision by selecting the services that best meet customer needs, in order to create a composite service.

1 INTRODUCTION

The company environment has strongly evolved in recent years; it has become more complex and unpredictable, moreover technological changes have greatly influenced the organization of the companies. The integration of these changes becomes essential to ensure companies' survival, not only at a single company, a sector or a given economy level but at global level (Grefen et al., 2009) companies are now trying to "squaring the circle" because they must firstly meet the competitive challenges by improving industrial performance in terms of cost, time, etc. Secondly answer the openness problems. Many companies have realized the importance of cooperation to ensure their survival. Cooperative spaces, in which companies work and react together have emerged in various forms: virtual company, companies' network etc. The establishment and the management of this type of organization are often based on technical and information sharing platforms which support and facilitate the cooperation (Grefen et al., 2009). We will focus in our work in this kind of relationship and more particularly to the notion of the on demand cooperation. We take the definition proposed by (Boukadi et al., 2010) of the on demand cooperation as a temporary grouping of partners distributed in space and time. This grouping is formed from opportunistic alliances initiated by a company called the initiator company of cooperation project which will be dissolved once the opportunity finished. Such a scenario involves the collaboration of different parts of a cooperative process that consists of several processes performed by different partners in order to meet a common purpose or to seize an opportunity in the market.

While the information technology innovates increasingly, but the industrial world is still reluctant to join dynamic cooperation scenarios, since the companies' information system is not suited to this type of operation.

The service-oriented architecture and Web service technology seem to provide credible answers to the needs both internally (information system support of the cooperation) and externally (services available to partners). But in the current state, several obstacles constrain the implementation of the SOA within the company as well as the forming of services-based cooperation. It is basically the lack of implementation method to define the architecture of services within the company information system. In addition, there are few works about the construction methods of the on-demand cooperative process based on services composition paradigm. To these

two problems a third one related to the dynamic nature of the cooperation is added. The analysis of the problem of services based inter-company cooperation has led us to the decomposition into three additional sub issues that shape our research problem, which are respectively the building problem of a new framework that ensures the effectiveness and efficiency of the intercompany cooperation based on service approach, the problem of decision making in order to choose the best services for a given query and the building problem of the on-demand collaborative process (or the composition of different services). Our contributions answer the limitations and problems previously described by offering mediation architecture between different companies which are a cooperation partners, the use of mediator ensures interoperability between companies and some level of security. To facilitate decision making and be able to choose the services that best meet customer needs, we propose the integration of decision trees in the software agents of our mediator. The third part of our contribution is the implementation of cooperation; this cooperation necessarily passes through an interconnection of processes of different companies, using the service composition paradigm.

This paper is structured as following: Section 2 and 3 are consecrated to a study of our proposal which will be illustrated by an example in section 4. Section 5 is dedicated to a comparative study between multiple research works finally in section 6, we will make conclusions about our work and we will talk about the extension of it.

2 THE PROPOSED MODEL

After having introduced the concept of Inter-enterprise cooperation that results in the formation of business process built from the interconnection of various processes and the partners have a partial master on the business process. We propose in this part mediation architecture between the services of the companies that are participating to the cooperation. According to J-L Lascoux (Touzi et al, 2008) the mediator "provides a framework with its own benchmarks, consisting of operating and communication rules, and steps process. It begins with the acknowledgement of the parts position in terms of legitimacy, to the formalization of an agreement as satisfactory as possible for the parts". In our architecture the mediator is used to perform a dynamic intercompany process. To ensure the interoperability of the participating companies we propose the integration of software agents in our mediator, learning of these agents is done by decision trees. The intercompany process (composite service created) will be published to be used as a web service, to permit the visibility and to be reused by other customers. Figure 1 shows an overview of our architecture.

Figure 1 shows our mediation architecture that consists of:

- Participating Companies: These companies offer accessible services via the public part of their information system. The Information about each service is stored in a services directory of the mediator.
- Mediator: is the core of the architecture, it is acting as an intermediate between the requester and the partners. It is composed of:

- Interface Agent: it receives and makes the processing of the applicant request to build a decision tree, which will be sent to the coordinator agent.
- Services Directory: at this directory we have services agents each of these agent is related to a service proposed by a company. These agents

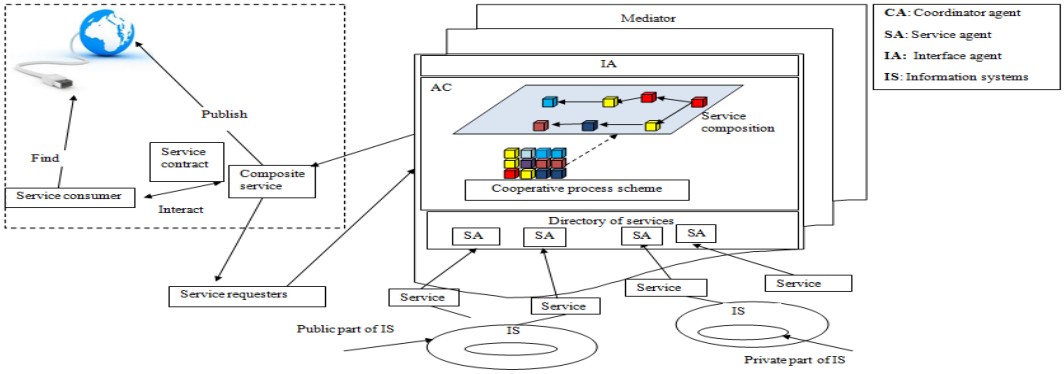


Figure 1: Architecture for intercompany cooperation.

represent information about each service, the information is organized as: a local decision tree in each agent service, and service address, communication protocol, and the details of the service purpose.

- Coordination Agent: it allows creating a composite service to meet the client request. It has a global decision tree that is created from the set of local decision trees. This tree is used to make a decision which allows choosing the best services to meet customer demand.

➢ Applicant service: it represents the launch process.

3 LEARNING AGENTS BY DECISION TREES

To enable the communication of different heterogeneous companies we propose to combine the software agents and decision trees.

Figure 2 shows the general structure of the coordination agent, respecting the three properties of independence, communication and intelligence. This architecture is inspired from the modular theory of (Fodor, 1983) that consists of five modules managing knowledge, perception, communication, control and the thinking of the agent.

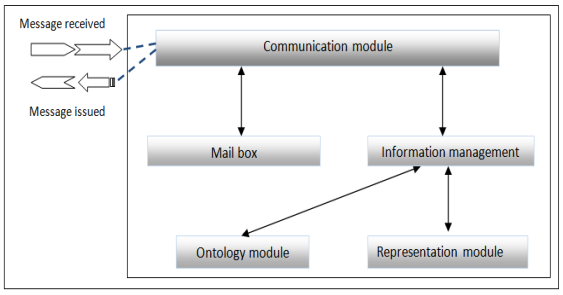


Figure 2: Components of the Coordinator Agent.

➢ Communication Module is a messages support between the agent and its coordinator. All interactions between the coordination agent and the services agent or the interface agent pass through this module.

➢ Mailbox is used to submit messages. This queue box type FIFO (First in First Out), is used for storing messages in order to process them asynchronously.

➢ Representation module receives the request sent by the agent service as a tree, and then reformulates the global tree (the tree that includes all local trees published by the different service agents) following the order of criteria send by interface agent

➢ Information Management Module contains information related to the service published by partners (protocol, address, message structure, description of the roles of services). Each agent service uses a local ontology to represent vocabulary and concepts in its application domain.

➢ Ontology module deals with the search for correspondences between agents to perform a cooperative task.

4 DECISION-MAKING MECHANISM BASED ON DECISION TREES

The coordinator agent uses a decision tree to quickly select the services that best meet customer needs, a decision tree can be used to clarify and find an answer to a complex problem. The structure allows our coordinator agent to take a problem with multiple possible solutions and display it in a simple and easily understandable format that shows the relationship between different events or decisions.

4.1 Structure of the Local Tree

The following figure shows the internal structure of our local trees:

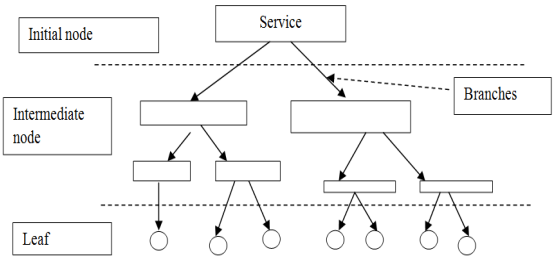


Figure 3: Internal structure of our local trees.

- *Initial Node:* each local tree initial node located in an agent service represents a company service.

-*Branches:* they connect an interior node" parent" to a "child" node. They represent the variable value tested in the parent intermediate node to the child node. Each branch corresponding to a question, allows choosing the path to follow.

Rule 1: the path to go in the tree: if the answer is yes, then the left path will be taken, otherwise it must move to the right one.

-*Intermediate Node:* represents the deals offered

by each service, each intermediate node corresponds to a response to determine the deal. Each answer then determines the next question.

Rule 2: each non-terminal node corresponds to a test $P(x)$ which evaluates the predicate P in Example x

-Leaf: each leaf refers to a class.

4.2 Training Algorithms by Decision Tree

The main training algorithms by decision trees are CART (Breiman et al, 1994) and C4.5 (Quinlan, 1984). We present below some of their characteristics. The first step is to build a small tree which consists of the most of data. The idea is to divide recursively and efficiently as possible the training sets examples by defined tests using attributes until we get subsets of examples which contain only examples belonging to the same class. This idea leads to a Top-Down construction methods that's mean constructing the tree from the root to the recursive leaves. We can then define a general scheme of algorithm.

Generic Training Algorithm:

```
BEGIN
  Initialize the current tree to the empty tree; the
  root is the current node
  REPEAT
    Decide whether the current node is terminal
    If the node is terminal then
      Assign a class to it
      Else
    Select a test and create as many new son nodes
    as there
    are possible answers to the test
      End if
    Pass to the next unexplored node if it exists until
    a decision tree is obtained
  End
END
```

5 EXAMPLE: EMERGENCY CASE

We will treat in this part the realization of a response process of an emergency case which is an accident, using three existing entities named: Hospital, SAMU and Police.

- The Hospital service: each hospital propose a description including its position, in our case we will need a GPS coordinates of the hospital.
- The SAMU service: provides a description of its services such as the emergency care team, the possible route and the expected duration of each route.

The Police service: provides the position of the police stations and available patrols.

- The description of each hospital the SAMU and The Police services is published in an agent service and represented by a local decision tree, and the coordination agents representing the global tree for each existing entities.

5.1 Operational Scenario

Using the three services mentioned previously, we will elaborate an emergency response scenario to carry out a sequence of actions for each accident report. We wish to create here a response emergency service which is called 333. This service will allow users to report an accident at a given position. The service will take the necessary steps and return to the user the expected intervention delay in seconds. The internal scenario which will lead the interactions between services is the following:

The coordinator agent will firstly ask the accident base to determine whether the reported accident is known, if so a negative value is simply returned to the service requester. Otherwise, we will first find the nearest hospital to the accident position, we use GPS data of the hospital and we determine both the SAMU and the police to be sent. We then add this to the accidents base before returning to the user the intervention time that was given by the SAMU.

5.2 Creation of Trees

We have chosen to use the SIPINA method for the generalization of our local trees, SIPINA is software but it is also a training method. It generalizes the trees by introducing an additional operation which is the fusion, during the induction of the prediction model. The figure 4 shows the generalization of the tree from Excel files, after following all method steps we obtain the following tree. After creating all local trees, the coordinator agent generates a global tree that includes all local trees; this tree that has been created manually, until the definition of an algorithm which allows the fusion of the local trees according to a test on each arc in order to create a tree that includes other trees by category.

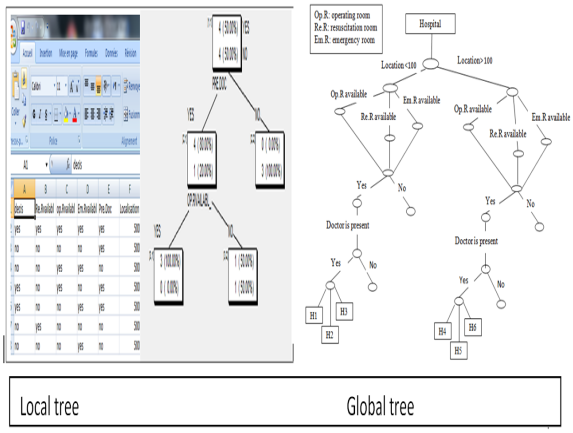


Figure 4: Generalization of the tree from Excel using Sipina Version 3.2.

5.3 Some Implementation Aspects

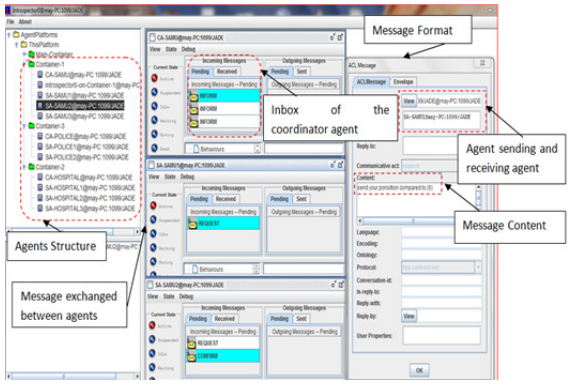


Figure 5: Communication between agents JADE.

Figure 5. shows the structure of our agents, and some messages exchanged between its, using the Jade platform to create the behavior of agents and to examine the feasibility of our proposition. For example coordinator agent of SAMU service named "CA-SAMU" sends a request to the service agents "SA-SAMU" the representatives of the different SAMU, then this services respond with a messages "inform" given their position, the coordinator agent selected the service that best meets the needs and sends a message "confirm" to the service agent concerned, the traces of the messages exchanged between agents are made with the help of sniffer agent. The third part of the figure shows the structure of a message "request" sent by the agent "CA-SAMU" to the "CA-SAMU" agent.

6 SOME RELATED WORK

Several models have been proposed to establish the enterprise collaboration. The proposed works have the information system at the center of their concerns and qualify it as an indispensable support for the implementation of the company strategy. The authors of some research work (Indulska et al, 2006) (Henderson et al, 2007) (Rolland et al, 2009) were attached to the analysis and design domain of process-oriented information systems. Several methods and tools of information systems engineering have emerged such as component-based approaches, MDA (Model Driven Engineering). These methods have been proposed to meet the evolution of the current information system context: evolution in terms of the information system architecture (heterogeneous and distributed), evolution in its use (open and accessible for different users). Since a few years the concept of Service Oriented architecture (SOA: Service Oriented Architecture) has been rapidly spread and widely accepted as a supporting architecture of the company information system. We studied the overall work according to the approaches and technologies used the nature of collaborative processes if it is static or dynamic, and the field possible for the use of the proposed solution.

(Touzi et al, 2008) adopt a non standardization approach, which proposes the design of a mediation system between information systems. The proposed practices are similar to the MDE and allow one to anticipate the tooling of a model transformation workshop. This method was also adopted in (Truptil et al, 2011), to solve a crisis management problem, the approach used in (Hui et al, 2010) is a standardization approach that consists on the proposition of a framework where all the participating company must respect it.

Among the works that have used the agent paradigm for solving the inter-company cooperation problems, the authors, (Brahimi et al, 2009) (Namin et al, 2006) have proposed global architectures, integrating the technology of web services. The difference that (Namin et al, 2006) proposed the integration of web services and software agents within the internal structure of the company and by adopting software agents within the UDDI registry; from the company side. Few studies have treated the notion of dynamic cooperation. For instance, the work of (Boukadi et al, 2010) offers such a mechanism of cooperation. (Mallek et al, 2011) have a different vision. They proposed an approach to detect problems before a real collaboration, then

analyzing and finding solutions to each partner. A few studies cited above that have used software agents and they have not addressed the decision-making of these agents. In this work we propose a distributed system, and when it comes to designing this type of system, agent technology is suitable, because MAS not only allow the sharing or distribution of knowledge, but also the achievement of a common goal. Agents in our work learn the decision from the user in certain situations to support him in other situations. The decision problem in the case of the cooperative distributed systems is a complex problem. Then, the training algorithm of the agent must meet several requirements: interactive, explicative, online, not handling only the quantitative but also the qualitative parameters. Also we must divide our decision problem into several problems. In our case, we choose to use decision trees, because they can model simply graphically and quickly a complex measured phenomenon. The aim of the use of training agents with the paradigm of decision trees is to select the best services that meet customer needs.

7 CONCLUSIONS

In this paper, we have proposed an architecture to create a dynamic process of cooperation; our architecture can be used in the case of B2B (business-to-business), B2C (Business-to-Consumer) or B2A (business-to- administration). This solution is based on approach of no-standardization; companies keep their structure and offer services. The mediator is involved in making the transformation between the services to create an intercompany collaborative process. We have integrated decision trees in a module of agents called representation module. The use of decision trees can reduce the response time to choose the best service. In order to better improve our approach, we will use the reinforcement learning (RL) algorithm, which has long been promising methods for enabling an autonomous agent to improve its behavior on sequential decision-making tasks. The next step is to specify a transformation model for integrating the decision trees in the agent modules.

REFERENCES

Boukadi, K., Vincent, L., Ghedira, C., Maamar, Z., 2010. CSMA: Context-Based, Service-Oriented Modeling and Analysis Method for Modern Enterprise Applications. In *International Journal of Organizational & collective Intelligence*, pp 1-28.

Brahimi, M., Boufaida, M., Seinturier, L., 2009 Multi-Agent Architecture for Developing Cooperative E-Business Applications. In *International Journal of Information Systems and Supply Chain Management (IJISSCM)*.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C., 1984. Classification and Regression Trees, Wadsworth. Inc., Belmont, California.

Fodor, J., 1983. *The Modularity of the Mind*, Cambridge, Mass, MIT Press.

Grefen, P., Kouvas, N., Weichhart, G., Eshuis, R., 2009. Dynamic business network process management in instant virtual enterprises, Computers inIndustry, pp. 86–103.

Henderson, S., Gonzalez, C-P., Ralyté, J., 2007 Situational method engineering: chunks or fragments. In the CAiSE ForumTrondheim, pp 89-92.

Hui, L., Anne, F. Cutting, D., Bourey, J.P., 2010. Use of Ontology for Solving Interoperability Problems between Enterprises. In the *11th IFIP Working Conference on virtual enterprises*.

Indulska, M., Green, P., Recker, J., Rosemann, M., 2006. Business Process Modeling: Perceived Benefits. In the *28th International Conference on Conceptual Modeling* Springer, Gramado, Brazil .

Mallek, S., Daclin, N., Chapurlat, N., 2011. An Approach for Interoperability Requirements Specification and Verification. In *Third International IFIP Working Conference*, pp 89-102.

Namin, A-S., Shen, W., Ghenniwa, H., 2006. Implementing enterprise collaboration using web services and software agents. In *International Journal of Nerworking and Virtual Organisations*, pp 185-201.

Quinlan, J. R., 1984. C4.5: Programs for Machine Learning, Morgan Kaufmann.

Rolland, C., 2009. Method engineering: towards methods as services, Software Process: *Improvement and Practice*, 14(3), pp. 143-164.

Touzi, J., Bénaben, F., Pingaud, H ., 2008. Prototype to Support Morphism between BPMN Collaborative Process Model and Collaborative SOA Architecture Model. In the *4th International Conference on Interoperability for Enterprise Software and Applications*, pp 145-157.

Truptil, S., Bénaben, F., Pingaud., P., 2011. On-the-fly adaptation of crisis response information system. In *Conference on Management of Emergent Digital EcoSystems*. pp114- 121.

The Effectiveness of Electronic Word of Mouth on Consumers' Perceptions of Adopting Products/Services
A Literature Review

Ali Reza Montazemi and Hamed Qahri Saremi
DeGroote School of Business, McMaster University, Hamilton, L8S 4M4, Ontario, Canada
{montazem, qahrish}@mcmaster.ca

Keywords: Electronic Word of Mouth, eWOM, Self-service Technology, SST, Review, Grounded Theory, Perception, Adoption Process, Social Communication Framework.

Abstract: The objective of this paper is to identify factors that influence the effectiveness of electronic word of mouth (eWOM) message on the consumers' perception of adopting products/services offered through self-service technology and brick-and-mortar channels. We draw on "social communication framework" and "Grounded Theory Literature Review Method" to systematically review the factors investigated in the extant eWOM literature to synthesize the effectiveness of eWOM message on consumers' perception of adopting products/services. This enabled us to identify several gaps in the body of knowledge in the eWOM literature that provide direction for future research to ameliorate the gaps.

1 INTRODUCTION

Consumers increasingly use online media channels (e.g., online discussion forums, consumer review sites, weblogs, social network sites) to share and use electronic word of mouth (eWOM) about products/services (Cheung and Thadani, 2012). eWOM is considered a powerful marketing force. For example, Facebook is selling user endorsement of products/services and Google is planning to do the same in the near future (Miller and Goel, 2013). The effectiveness of eWOM on consumers' perceptions of adopting products/services has been studied extensively in recent years (Cheung and Thadani, 2012; Wu and Gaytan, 2013). However, the scope of the published studies in this area is fragmented (Cheung and Thadani, 2012), making it difficult to draw systematic conclusions about the effectiveness of eWOM. As will be shown in this paper, researchers have identified 173 constructs within three contexts (i.e., self-service technology (SST), brick-and-mortar, and unknown) and three stages of the adoption process (i.e., pre-usage, initial use, and repeated use) to investigate the effectiveness of eWOM on the consumers' perception of adopting products/services.

Adoption of products/services by consumers involves a process that occurs over time and consists of three stages, with actions and decisions occurring at each stage (Rogers, 2003). Pre-usage stage of adoption process begins with awareness that leads to mental evaluation of the products/services, which in turn may lead to consumers' initial use of the products/services (Looney et al., 2008). Initial use consists of trial that may lead to consumers' repeated use of the products/services (Looney et al., 2008).

This paper presents a systematic review of the extant eWOM literature to synthesize the factors that shape the effectiveness of eWOM on the consumers' perception of adopting products/services, across three stages of adoption process. Our findings identify critical knowledge gaps in need of future investigation. To save space, we use eWOM to represent "eWOM message" and adoption process to represent three stages of "pre-usage, initial-use, and repeated-use".

This paper is organized as follows. First, we present a theoretical framework for synthesizing the extant literature on eWOM. Next, based on the five steps of "Grounded Theory Literature Review Method", (Wolfswinkel et al., 2013), we define our methodology for analyses of the extant eWOM literature. Section 4 provides a comprehensive summary of our findings and a classification of the factors that shape the effectiveness of eWOM on the consumers' perception of adopting products/services

offered through both SSTs and brick-and-mortar channels. We conclude by discussing the implications of our findings for further theoretical and empirical investigations.

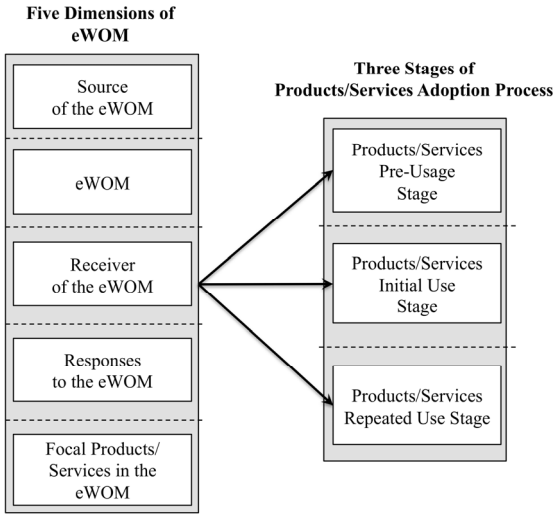


Figure 1: Our Conceptual Framework.

2 THEORETICAL FOUNDATION

eWOM is considered as a prominent form of social influence that affects the consumers' perceptions of products/services adoption process (Bock et al., 2012). Social influence is the extent to which members of a social network influence one another's perceptions. This influence is exerted through messages (signals) that help to form perception of the value of a product/service, at each stage of the adoption process (Montazemi and Qahri Saremi, 2013). eWOM, as a form of social influence, is consumer-generated product/service evaluations posted on the company or a third party websites (Mudambi and Schuff, 2010). The rise of online media channels during the last two decades has offered fertile ground for eWOM to become an important means of influencing consumers' perception of pertinent product/services offered through self-service technologies (SSTs) and/or brick-and-mortar companies (Van Hoye and Lievens, 2007).

We draw on social communication framework (Hovland, 1948) as the theoretical underpinning to guide our literature analysis. Social communication is defined as "the process by which an individual (the communicator) transmits stimuli to modify the behavior of other individuals (receivers)" (Hovland, 1948, p. 371). The eWOM, as a form of online social communication media, comprises five major dimensions that are depicted in Figure 1: (1) the eWOM communicator (source) who transmits the eWOM, (2) the eWOM as a form of social communication content (stimulus) transmitted by the source, (3) the consumer who receives and responds to the eWOM (receiver), (4) the responses made to the eWOM by the receiver, and (5) the focal products/services in the eWOM. We systematically analyze eWOM literature to identify the factors representing each of the five dimensions of eWOM, presented next.

3 METHODOLOGY

We adopted the "Grounded Theory Literature Review Method" (Wolfswinkel et al., 2013), to review the extant eWOM literature. This method enabled us to perform a thorough and theoretically relevant analysis of the eWOM literature. The "Grounded Theory Literature Review Method" is implemented in five steps, depicted in Figure 2. Step 1 (called "Define") consists of: defining the inclusion/exclusion criteria, determining the pertinent databases of the published studies for search, and determining the appropriate search terms. In step 2 (called "Search"), the actual search for the published studies is performed using the keywords and the pertinent databases determined in the step 1. In step 3 (called "Select"), we refine the retrieved studies from step 2 using the inclusion/exclusion criteria that were determined in

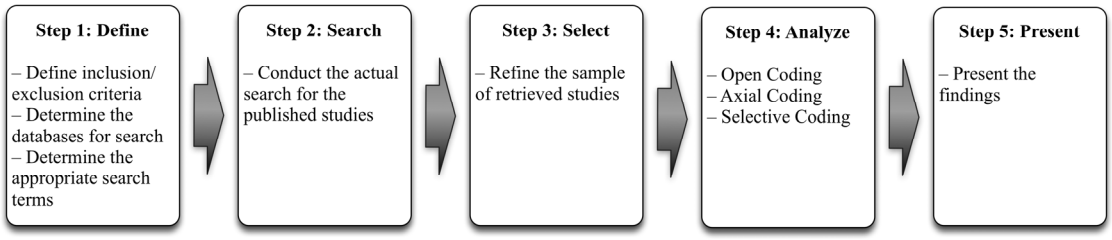


Figure 2: Grounded Theory Literature Review Method (Adapted from Wolfswinkel et al., 2013).

step 1. Next, in step 4 (called “Analyze”), we synthesize the refined sample of pertinent studies using Grounded Theory techniques. In step 5 (called "Present"), we present the findings and insights gleaned from the synthesis of the pertinent studies in step 4. Detailed description of the five steps in support of our syntheses of extant literature on eWOM is presented next.

3.1 Steps 1 – 3: Define, Search, and Select

To identify journal and conference papers that investigated the effectiveness of eWOM on the consumers' perception of products/services adoption process, we searched a number of electronic databases such as AIS, ACM, ScienceDirect, Palgrave Macmillan, Sage, EBSCOhost, JSTOR, Scholar’s Portal, and Google Scholar.

We used search terms of "word-of-mouth", "electronic word-of-mouth", "WOM", "eWOM", "online reviews", and "online recommendations". To ensure that no major articles is ignored, we also searched the leading scholarly IS journals (i.e., MIS Quarterly, Information Systems Research, Journal of Management Information Systems, European Journal of Information Systems, Information Systems Journal, Journal of Strategic Information Systems, Journal of Association for Information Systems, and Information & Management), and four major scholarly Marketing journals (Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research, and Journal of the Academy of Marketing Science). The search initially yielded 181 studies. The nature of empirical environment for these 181 studies can be categorized in terms of three contexts: SST, brick-and-mortar, and unknown. SST represents studies that were conducted within the context of products/services in an online environment (i.e., click-and-mortar and/or pure play). Brick-and-mortar studies assessed the effectiveness of eWOM on consumers’ perception of adopting products/services in a traditional offline environment. Finally, the “unknown” context represents those investigations that did not report the nature of the environment for their study.

Not all the 181 studies were appropriate for inclusion in our analysis. Recognizing this, Wolfswinkel et al., (2013) recommend that researchers mark out the scope of their review by establishing criteria for inclusion. To that end, we included only peer-reviewed studies in which (1) eWOM was the main focus of investigation in the paper, and (2) the research was empirical. Applying these two criteria to 181 papers resulted in 47 eWOM studies that we used in our analysis and can be obtained from:
http://profs.degroote.mcmaster.ca/ads/montazem/papers/WEBSIT_paper_suppliment.pdf

3.2 Step 4: Analyze

To analyze the selected 47 eWOM studies, constructs used in those studies were initially identified and coded. Next, we synthesized them into factors based on their conceptual similarity and map them on the five dimensions of eWOM, using three Grounded Theory techniques: (1) *open coding*, (2) *axial coding*, and (3) *selective coding*, which are explained next.

We used *open coding* technique, through which constructs, their stage (i.e., pre-usage, initial use, and repeated use), and context (i.e., SST, brick-and-mortar, and unknown) of each study were coded. We coded constructs based on their instrument measures used in the eWOM study, rather than on the author’s label because the labels that authors applied to their constructs can vary (Montazemi et al., 2012, Montazemi and Qahri Saremi, 2013). The outcome was 173 constructs used in the 47 eWOM studies. Next, we applied *axial coding* technique, to synthesize these 173 constructs based on their conceptual similarities that resulted in 30 factors. *Selective coding* technique was used to map these 30 factors on the five dimensions of eWOM, depicted in Figure 3. Next, we present our findings.

4 RESULTS (STEP 5: PRESENT)

Due to the space limitation, definitions and references for each factor are not reported in the Tables 1 to 5 that appear in this section. These details can be obtained from:
http://profs.degroote.mcmaster.ca/ads/montazem/papers/WEBSIT_paper_suppliment.pdf.

4.1 Factors Representing the Response to eWOM

We identified twelve factors that represent the “response” dimension of effectiveness of eWOM, depicted in Table 1.

Six of these factors (i.e., adoption of eWOM, purchase/use intention, perceived helpfulness of eWOM, attitude towards the product/service, trust in the vendor, and perceived credibility of eWOM) have been studied across the three stages of the

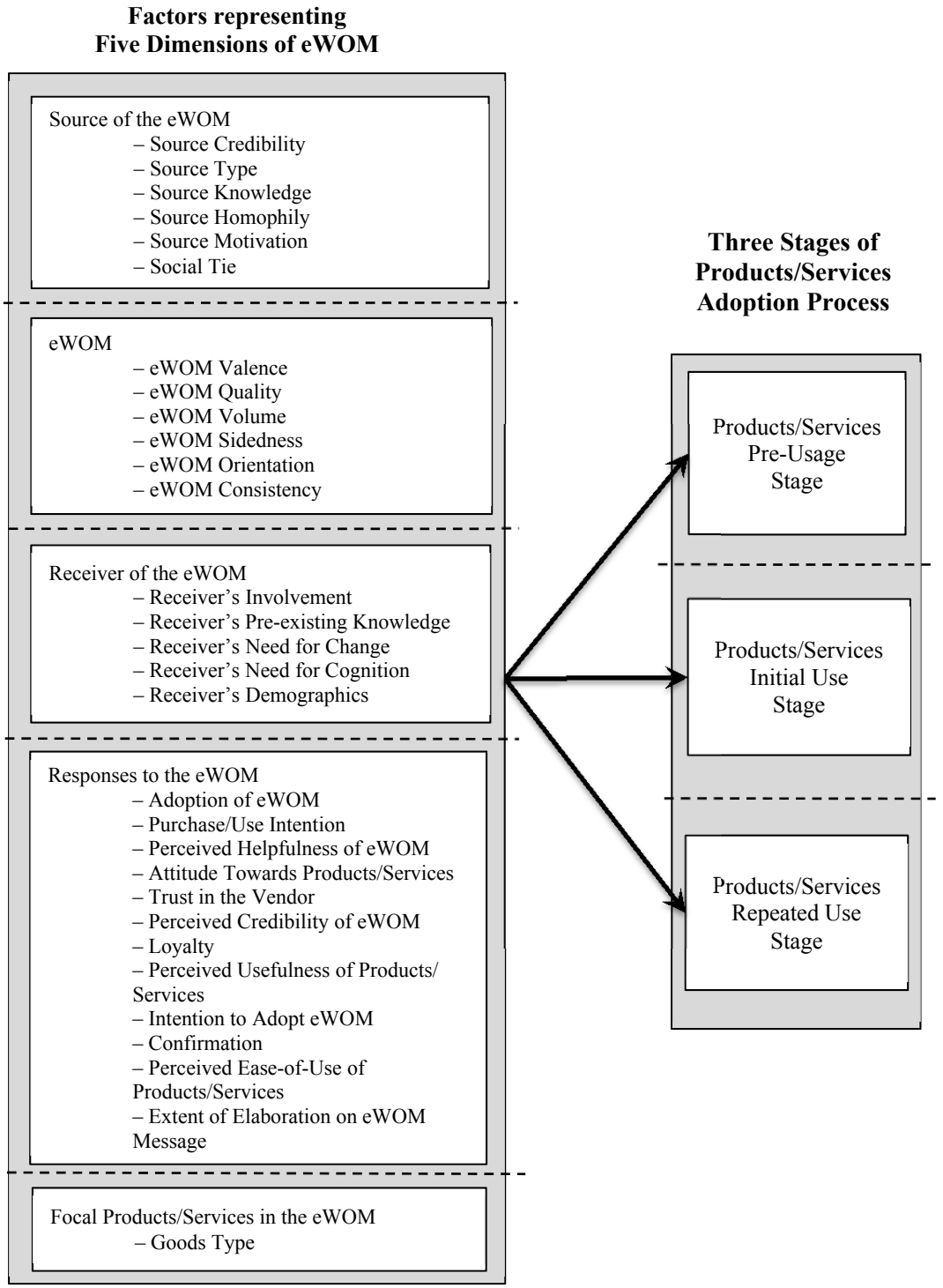


Figure 3: The 30 Factors representing Five Dimensions of eWOM in the proposed Conceptual Framework.

adoption process, in the SST context. However, there are gaps in the body of knowledge in the eWOM literature for the SST context in regard to the remaining 6 factors. For instance, effectiveness of eWOM on consumers' trust in vendor, which is an important factor in forming consumers' perception of SST adoption (Bock et al., 2012), differ significantly across stages of adoption process.

Table 1: Number, stage, and the context of the eWOM studies investigating each factor representing response to eWOM.

Factors Representing response	Context	P*	I*	R*
Adoption of eWOM	SST*	1	1	3
	B&M*	2		2
	NA*	4		1
Purchase/ Use intention	SST	7	2	2
	B&M	2		
	NA			
Perceived Helpfulness of eWOM	SST	2	3	4
	B&M			1
	NA	1		
Attitude towards the product/service	SST	1	2	2
	B&M	2		
	NA	1		1
Trust in the Vendor	SST	3	3	3
	B&M			
	NA			
Perceived Credibility of eWOM	SST	3	2	3
	B&M			
	NA			1
Loyalty	SST			
	B&M	1		1
	NA	2		1
Perceived Usefulness of the product/service	SST		1	1
	B&M	1		
	NA			
Intention to Adopt eWOM	SST		2	
	B&M			1
	NA			
Confirmation	SST			
	B&M			1
	NA			1
Perceived Ease of Use of the product/service	SST			1
	B&M			
	NA			
Extent of Elaboration on eWOM Message	SST			
	B&M			
	NA	1		

* Note:
SST: Self-Service Technology Context
B&M: Bricks-and-Mortar Context
NA: Context Unknown
P: Pre-usage Stage
I: Initial-use Stage
R: Repeated-use Stage

Nonetheless, there is a gap in the literature whether the same pattern exists for other factors of consumers' perception such as consumers' perceived ease-of-use and perceived usefulness of an SST. Furthermore, as indicated in Table 1, none of the twelve factors have been studied across the three stages of the adoption process, in the bricks-and-mortar context. For instance, although we know that effectiveness of eWOM on consumers' trust in vendor differs significantly across stages of adoption process, in the SST context, we do not know if the same pattern exists for the effectiveness of eWOM on consumers' trust in vendor in the bricks-and-mortar context. Such a fragmented scope of studies in the eWOM literature makes it infeasible to draw systematic conclusions about the effectiveness of eWOM on consumers' perception of adopting products/services.

4.2 Factors Representing the eWOM Message

We identified six factors representing the eWOM that have been investigated in the eWOM studies, depicted in Table 2. One of these factors, eWOM valence, has been studied across the three stages of SST adoption process. However, there are gaps in the body of knowledge in regard to all of the six factors across the three stages of bricks-and-mortar adoption process. For instance, the literature shows that the effect of eWOM valence on consumers' perceived usefulness of an SST differs significantly

Table 2: Number, stage, and the context of the eWOM studies investigating each factor representing eWOM message.

Factors Representing eWOM Message	Context	P*	I*	R*
eWOM Valence	SST*	4	2	5
	B&M*	4		1
	NA*	3		1
eWOM Quality	SST	5		1
	B&M			1
	NA	2		2
eWOM Volume	SST	5	1	
	B&M			
	NA			
eWOM Sidedness	SST			
	B&M	1		
	NA			1
eWOM Orientation	SST	1		
	B&M			
	NA	1		
eWOM Consistency	SST			
	B&M			
	NA			1

* Note:
SST: Self-Service Technology Context
B&M: Bricks-and-Mortar Context
NA: Context Unknown
P: Pre-usage Stage
I: Initial-use Stage
R: Repeated-use Stage

across different stages of SST adoption process (Vermeulen and Seegers, 2009). Nonetheless, there is a gap in the literature whether the same pattern exists for the bricks-and-mortar context.

4.3 Factors Representing the Source of the eWOM

We identified six factors representing the source of eWOM, depicted in Table 3. None of these factors have been studied across the three stages of adoption process in the SST and bricks-and-mortar contexts. This represents gaps in the body of knowledge in the eWOM literature as to how these six factors affect the effectiveness of eWOM on consumers' perception of adopting products/services. As a case in point, consumers read eWOM about the products/services from a vast and geographically dispersed group of strangers through blogs, online consumer review platforms, shopping bot sites, and social networking sites (Cheung and Thadani, 2012). This has the potential to raise consumers' concern about the credibility of the source of eWOM. In such an uncertain situation, high degree of consumers' experience with the products/services is expected to have a significant positive effect on consumers' perception of adopting products/services. However, eWOM literature is silent on this issue within the context of SST and bricks-and-mortar.

Table 3: Number, stage, and the context of the eWOM studies investigating each factor, representing source of the eWOM.

Factors Representing source of the eWOM	Context	P*	I*	R*
Source Credibility	SST*	2		1
	B&M*			1
	NA*	3		2
Source Type	SST	3	2	
	B&M	1		
	NA	1		
Source knowledge	SST			
	B&M	1		3
	NA			1
Source Homophily	SST			
	B&M	1		1
	NA			
Source Motivation	SST			
	B&M			1
	NA			1
Social Tie	SST			
	B&M	1		
	NA			

* Note:
SST: Self-Service Technology Context
B&M: Bricks-and-Mortar Context
NA: Context Unknown
P: Pre-usage Stage
I: Initial-use Stage
R: Repeated-use Stage

4.4 Factors Representing the Receiver of the eWOM

We identified five factors representing the receiver of the eWOM, depicted in Table 4. None of them have been studied across the three stages of adoption process. This represents a gap in the body of knowledge in the eWOM literature as to how these five factors affect the effectiveness of eWOM on consumers' perception of adopting products/services. For example, the eWOM literature provides support for the importance of consumers' pre-existing knowledge, need for cognition, and need for change on the effectiveness of eWOM on the consumers' perception of adopting products/services. Nonetheless, there is a gap in the eWOM literature whether these factors have the same effects across the three stages of the adoption process. Considering the significance of the stage of the adoption process in consumers' perceptions of

Table 4: Number, stage, and the context of the eWOM studies investigating each factor, representing receiver of the eWOM.

Factors Representing receiver of the eWOM	Context	P*	I*	R*
Receiver's Involvement	SST*	3		1
	B&M*			
	NA*	1		2
Receiver's Pre-existing Knowledge	SST	1		2
	B&M	1		1
	NA	1		2
Receiver's Need for Change	SST			
	B&M			
	NA	1		
Receiver's Need for Cognition	SST		1	
	B&M			
	NA			
Receiver's Demographics	SST			1
	B&M			
	NA			

* Note:
SST: Self-Service Technology Context
B&M: Bricks-and-Mortar Context
NA: Context Unknown
P: Pre-usage Stage
I: Initial-use Stage
R: Repeated-use Stage

adopting products/services, such a lack of knowledge represents a gap in the eWOM literature.

4.5 Factors Representing the Focal Product/Service in the eWOM

A key determinant of eWOM effectiveness on consumers' perception of adopting products/services is the nature of the product/service (i.e., goods type) under consideration, as indicated in Table 5. According to Nelson (Nelson, 1970), *search goods* are the products/service for which consumers have the ability to obtain information on their quality prior to adoption, while *experience goods* are products/services that require sampling or adoption to evaluate their quality. Examples of search goods include cameras and natural supplement pills (Mudambi and Schuff, 2010). Examples of experience goods include music (Bhattacharjee et al., 2006) and wine (Mudambi and Schuff, 2010). Although the eWOM literature indicates that goods type influences the effectiveness of eWOM on the consumers' perception of adopting products/services in the SST context (indicated in Table 5), we do not know if the same pattern exists in the context of bricks-and-mortar. Considering the importance of the goods type in consumers' perceptions of adopting products/services, such a lack of knowledge represents a gap in the eWOM literature.

Table 5: Number, stage, and the context of the eWOM studies investigating each factor, representing focal product/service in the eWOM.

Factors Representing the focal product/service in the eWOM	Context	P*	I*	R*
Goods Type	SST*	2	2	1
	B&M*			
	NA*	1		1

* Note:
SST: Self-Service Technology Context
B&M: Bricks-and-Mortar Context
NA: Context Unknown
P: Pre-usage Stage
I: Initial-use Stage
R: Repeated-use Stage

5 DISCUSSION

This paper tackles the question of how eWOM affects consumers' perceptions of adopting the products/services through a literature review based on a new theoretical framework. To that end, we drew on grounded theory literature review method and social communication framework to identify and systematically review 47 pertinent eWOM empirical studies. We have synthesized 30 conceptually distinct factors representing the five dimensions of the eWOM. Based on our findings, we identified numerous important gaps in the body of knowledge in the eWOM literature, which are in need of attention in the future research. We find that despite more than a decade of research, the extant eWOM literature remains largely fragmented. Researchers have chosen to study factors of their individual interest and there is little systematic integration among them. As a result, we observe findings that are inconsistent or even contradictory. One example is how the relationship between consumers' prior knowledge of the product/service and the effectiveness of the eWOM on the consumers' perceptions of adopting the product/service is conceived. On the one hand, Chatterjee (2001), Chakravarty et al. (2010), and Fan and Miao (2012) identify a negative relationship between the consumer's prior knowledge of the product/service and the effectiveness of eWOM on the consumer's perceptions of adopting the product/service. On the other hand, Park and Lee (2009), Park and Kim (2008), and Cheung et al. (2009) show a positive relationship between the two factors. Such a contradictory findings point to possible missing moderators such as need for cognition and need for change. Prior studies (e.g., Gupta and Harris, 2010, Wood and Swait, 2002) have shown that consumers with different levels of need for cognition and need for change, ceteris paribus, behave in systematically different fashions regarding their motivation to seek and process information about the products/services. Furthermore, scholars (e.g., Sussman and Siegal, 2003) have identified the moderating effect of consumer's motivation to seek and process a message about a product/service on the relationship between the consumers' prior knowledge of the product/service and the effectiveness of the message on the consumers' perceptions of adopting the product/service. Notwithstanding their importance on the effectiveness of eWOM message, the literature is silent on the moderating effects of consumers' need for cognition and need for change in all three stages of the products/services adoption process.

REFERENCES

Bhattacharjee, S., Gopal, R. D., Lertwachara, K. & Marsden, J. R. 2006. Consumer search and retailer strategies in the presence of online music sharing. *Journal of Management Information Systems,* 23**,** 129-159.

Bock, G. W., Lee, J., Kuan, H. H. & Kim, J. H. 2012. The Progression of Online Trust in the Multi-Channel Retailer Context and the Role of Product Uncertainty. *Decision Support Systems,* 53**,** 97-107.

Chakravarty, A., Liu, Y. & Mazumdar, T. 2010. The Differential Effects of Online Word-of-Mouth and Critics' Reviews on Pre-release Movie Evaluation. *Journal of Interactive Marketing,* 24**,** 185-197.

Chatterjee, P. 2001. Online reviews: do consumers use them? *Advances in consumer research,* 28**,** 129-133.

Cheung, C. M. K. & Thadani, D. R. 2012. The Impact of Electronic Word-of-Mouth Communication: A Literature Analysis and Integrative Model. *Decision Support Systems,* 54**,** 461–470.

Cheung, M. Y., Luo, C., Sia, C. L. & Chen, H. 2009. Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce,* 13**,** 9-38.

Fan, Y.-W. & Miao, Y.-F. 2012. Effect of electronic word-of-mouth on consumer purchase intention: The perspective of gender differences. *International Journal of Electronic BusinessManagement,* 10**,** 175-181.

Gupta, P. & Harris, J. 2010. How e-WOM recommendations influence product consideration and quality of choice: A motivation to process information perspective. *Journal of Business Research,* 63**,** 1041-1049.

Hovland, C. I. 1948. Social communication. *Proceedings of the American Philosophical Society,* 92**,** 371-375.

Looney, C. A., Akbulut, A. Y. & Poston, R. S. 2008. Understanding the Determinants of Service Channel Preference in the Early Stages of Adoption: A Social Cognitive Perspective on Online Brokerage Services. *Decision Sciences,* 39**,** 821-857.

Miller, C. C. & Goel, V. 2013. Google to Sell Users' Endorsements. *The New York Times*, October 12, 2013.

Montazemi, A. R., Pittaway, J. J., Qahri Saremi, H. & Wei, Y. 2012. Factors of stickiness in transfers of know-how between MNC units. *The Journal of Strategic Information Systems,* 21**,** 31-57.

Montazemi, A. R. & Qahri Saremi, H. Factors Affecting Internet Banking Pre-Usage Expectation Formation. *Hawaii International Conference on Systems Sciences (HICSS 2013)*, 2013 Maui, Hawaii. IEEE.

Mudambi, S. M. & Schuff, D. 2010. What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS Quarterly,* 34**,** 185-200.

Nelson, P. 1970. Information and consumer behavior. *The Journal of Political Economy,* 78**,** 311-329.

Park, C. & Lee, T. M. 2009. Information direction, website reputation and eWOM effect: A moderating role of product type. *Journal of Business Research,* 62**,** 61-67.

Park, D. H. & Kim, S. 2008. The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications,* 7**,** 399-410.

Rogers, E. M. 2003. *Diffusion of Innovations,* New York, The Free Press.

Sussman, S. W. & Siegal, W. S. 2003. Informational influence in organizations: An integrated approach to knowledge adoption. *Information Systems Research,* 14**,** 47-65.

Van Hoye, G. & Lievens, F. 2007. Social Influences on Organizational Attractiveness: Investigating If and When Word of Mouth Matters. *Journal of Applied Social Psychology,* 37**,** 2024-2047.

Vermeulen, I. E. & Seegers, D. 2009. Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management,* 30**,** 123-127.

Wolfswinkel, J. F., Furtmueller, E. & Wilderom, C. P. M. 2013. Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems,* 22**,** 45–55.

Wood, S. L. & Swait, J. 2002. Psychological indicators of innovation adoption: Cross-classification based need for cognition and need for change. *Journal of Consumer Psychology,* 12**,** 1-13.

Wu, J. & Gaytan, E. A. A. 2013. The role of online seller reviews and product price on buyers' willingness-to-pay: a risk perspective. *European Journal of Information Systems,* 22**,** 416-433.

Communities of Practice Supporting Research and Technological Development

A Practical Case at Electrical Sector in Mexico

Yasmín Hernández[1], Salvador González-Castro[2] and G. Arroyo-Figueroa[1]
[1]*Gerencia de Tecnologías de la Información, Instituto de Investigaciones Eléctricas, Reforma 113, Cuernavaca, Mexico*
[2]*División de Tecnologías Habilitadoras, Instituto de Investigaciones Eléctricas, Reforma 113, Cuernavaca, Mexico*
{myhp, sgc, garroyo}@iie.org.mx

Keywords: Communities of Practice, Innovation, Knowledge Management, Research, Technological Development.

Abstract: The IIE is devoted to innovation, technological development and applied scientific research. In order to increase productivity and to anticipate the needs of electrical sector, IIE is promoting collaboration and sharing of knowledge, ideas and experience among researchers through communities of practice. Communities of practice enable creating and sharing of knowledge and its application in the development of research projects, realize and anticipate problems, and devised new research lines to be undertaken. To impulse communities of practice, several efforts have been conducted, such as funding, talks by experts, defining success indexes and regulations and the development of an information system based on Web technologies. The Web system supports distance collaboration and sharing of documents. This paper presents an approach to implement communities of practice.

1 INTRODUCTION

In the global context, companies require endlessly for new knowledge in order to be innovative and competitive. Usually, this new knowledge comes from academic education and training of personnel; however, companies need to learn from their own experience. They also need means to manage their own knowledge, this includes means to generate, share and transfer it. The Electrical Research Institute (IIE, for the Spanish *Instituto de Investigaciones Eléctricas*) is a public company devoted to innovation, technological development and applied scientific research, in order to develop technologies for the electrical and oil industries. Its mission is to promote and support innovation through value-added applied research and technological development (IIE, 2013).

IIE's researchers are specialized in several fields of sciences and technology. They develop high quality projects. Most of these projects are multi-disciplinary and include researchers from different departments. IIE identified that workgroups work isolated often without noticing what other workgroups are doing; sometimes trying to solve similar problems. To promote knowledge sharing a knowledge management model is being implemented. Through several strategies, this model enables to generate, identify and share resources of knowledge across departments and divisions. Communities of practice are an important knowledge management strategy which contributes to organization's strategic direction; they enable expertise to be transferred across organization and encourage the discussion of effective solutions to a range of problems (Wenger and Snyder, 2000).

At IIE, communities of practice are groups of researchers who meet regularly to share specialized knowledge and learn together. They apply this knowledge in the development of innovative projects, products and services to be delivered to customers.

Formerly, informal groups of researchers met up to talk about an interesting paper or final outcomes of a project for instance. However, researchers did not collaborate, and it was difficult to know if researchers applied this knowledge in their projects. These groups have received recognition and resources to create formal groups with the aim to produce knowledge and results aligned with strategic vision. Provided resources includes a Web system to support collaboration.

Communities of practice has been identified as a key to improving performance, therefore a growing number of people and organizations in various sectors are now focusing on them (Wenger, McDermott and Snyder, 2002).

This paper presents the efforts in promoting communities of practices in the IIE. The rest of this section presents a brief description of the context of communities of practice. Sections 2 and 3 describe how communities of practice are integrated and how they work. In section 4, a Web system for communities of practice is described. Section 5 outlines how we are planning to measure communities of practice's success. Finally, conclusions and future work are presented in Section 6.

1.2 Knowledge Management

The knowledge management model is based on a broadly accepted definition, it points out that knowledge management is the process of managing the corporation's knowledge through a systematically and organizationally specified process for acquiring, organizing, sustaining, applying, sharing and renewing both the tacit and explicit knowledge of employees to enhance organizational performance and create value (Davenport and Prusak, 2000).

The model is intended to increase productivity and competitiveness through collaboration, and turns around seven supports: communities of practice, technological repository, relational capital, innovation management, intellectual property, competence management and human capital (IIE, 2013). Figure 1 shows the knowledge management model.



Figure 1: Knowledge management model.

The main domains of the system of knowledge are lined up with the technological development areas of the IIE, which in turn are aligned with the strategic planning of electrical and energy sectors in México. The principal customer of IIE is the Federal Electricity Commission (CFE) which is a company created and owned by the Mexican government. It generates and distributes electric power for almost 35.3 million customers what represents almost 100 million people (CFE, 2013). Thus, the domains of knowledge support electricity generation processes, such as: energy efficiency and energy savings, use of renewable energy sources, use reduction of fossil fuels and climate change mitigation.

2 COMMUNITIES OF PRACTICE

Communities of practice are a fundamental component of knowledge management because they join people to share knowledge and learn together about common topics of interest; they manifest the importance of social interaction in learning. The term "communities of practice" was coined by Etienne Wenger and Jean Lave (Lave & Wenger, 1991) and it is defined as groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an on-going basis (Wenger, McDermott and Snyder, 2002).

A community of practice (CoP) has three essential components: domain, community and practice (Wenger, McDermott and Snyder, 2002). *Domain* is the definition of the technological or research area of shared interest. In the IIE, *domain* corresponds or is lined up with the technological development lines of the IIE. The *community* is the group of researchers with a sense of belonging to it. CoPs are shaped across the organizational structure; therefore, CoPs are multidisciplinary and interdepartmental groups. The *practice* is given in terms of projects development which in turn originates a body of knowledge, methods, stories, cases, tools, documents, etc. The body of knowledge is produced and applied in the development of projects, which mainly consist in applied research and technological development. Figure 2 shows these components.

We have adopted a bounded definition: a CoP is a group of researchers constituted to develop specialized knowledge on a strategic theme based on shared learning and joint reflection about practical experiences (Hernández, Arroyo-Figueroa and González-Castro, 2013).

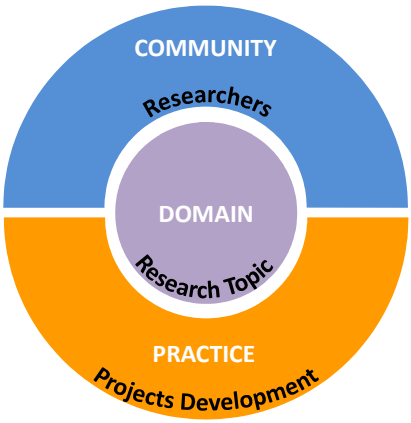


Figure 2: CoP components.

In some way, CoPs are not new in the IIE, since along its 35 years of live, several homogenous groups of researches have regular meeting to talk about several problems or to share knowledge and experience. These informal meetings were plenty of knowledge but most times, the knowledge was only in the verbal channel, but there was no history on how a problem was solved or how the knowledge was applied. On the other hand, researchers worked isolated in their projects and the knowledge and experience was no shared, even there were other projects facing similar problems. Moreover, sometimes research projects were no line up with strategic topics. Consequently, it was noticed the need to capture or retain that knowledge and to focus on topics around strategic lines with the aim to innovate.

As a result of these needs and the strategic planning, it was launched the innovation development programme (PDI, for the Spanish *Programas de Desarrollo de Innovación*). PDI aimed at the development of lines of research to support the competitiveness. A PDI involved a set of elements, such as projects, funding, laboratories, researchers, students, potential customers, and so on which are intended to support the development of a research line.

PDIs made evident the need for mechanisms to establish guidelines and responsibilities for knowledge management and innovation. As a result, communities of practice were initiated, as groups of researchers to develop specialized knowledge on strategic topics.

Communities of practice are formed with the purpose of working on strategic topics associated with lines of technological development. CoPs Members share experiences and explore new technologies to detect opportunities, where IIE, its customers and potential customers, may have a competitive advantage. In this way, CoPs support institution and its customers' competitiveness; at the same time CoPs supports professional development of members of CoPs. To achieve the objectives, communities of practice share knowledge, experiences and learning based on joint reflection.

The process to implement and shape CoP started with strategic goals statements and required core competencies, business processes, and key activities. These requirements were analyzed in terms of critical knowledge domains. Then potential CoPs members were identified. A requirement is people's interest and need for this knowledge for their work. Also the passion and attitude is important for the success of CoP. Finally, it was analyzed how to connect them into communities of practice in such a way they can share and generate knowledge. A CoP creation is strongly supported in energy sector and IIE strategic planning. Figure 3 shows the process to shape a CoP.

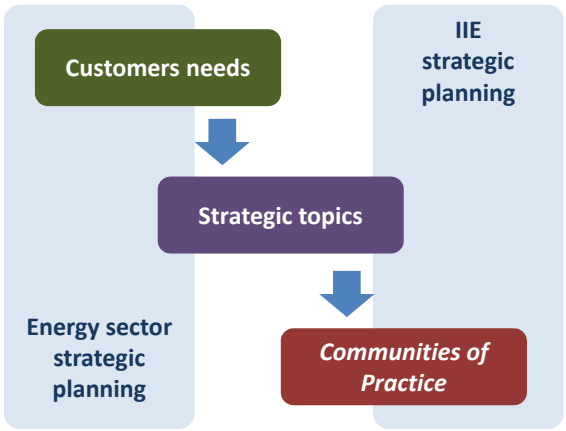


Figure 3: CoP process.

As we mentioned, a frequent problem was researchers worked alone and regularly the knowledge was not shared or reused; and, most times when a project finished, projects team split. Therefore there was necessity for CoPs composed by researchers from several departments and divisions participating in different projects, but with a permanent relationship in the CoP. Figure 4 explains this organization, in a schematic way.

In this way, a community of practice is a group of researchers constituted to develop specialized knowledge on a strategic topic. Currently, in the Enabling Technologies Division there are about 12 communities of practice and more than 100 researchers participating. Table 1 presents a list of CoPs.

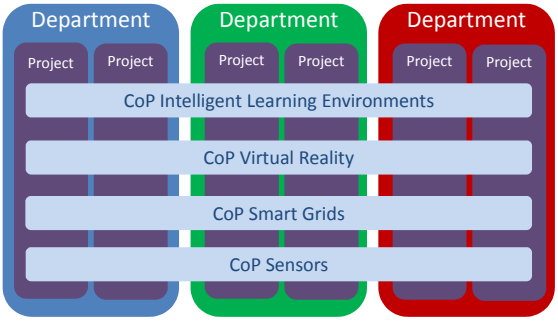


Figure 4: Horizontal organizational structure of CoP.

Table 1: List of communities of practice.

	Communities of practice
1	Systems for detection, diagnosis, prediction and remediation faults
2	Development of real and virtual sensors
3	Network and communication devices for process and elements monitoring
4	Smart grids
5	Virtual reality
6	Intelligent learning environments
7	Technologies for training simulators
8	Generation substation automation
9	Graphical interfaces
10	Business intelligence
11	Mechatronics
12	Power electronics

3 CoP OPERATION

To foster communities of practice, we have followed the guidelines proposed by Wenger (Wenger, McDermott and Snyder, 2002) but we have put some formal mechanism to monitoring CoP. We have a process document and a regulations document.

The process document establishes guidelines to create, update and cancel communities of practice. The regulations document establishes a minimum of guidelines for operation and monitoring of communities of practice.

CoPs can be set up with researchers who are interested in the domain of CoP. Preferably the members of a CoP must be knowledgeable in the subject matter and be experienced to support the work of the CoP, or to have the potential to get professionally developed in the subject domain. Communities of practices are constituted by a coordinator, a secretary and members. All of them must have the above characteristics.

To achieve goals, CoPs have some functions to be conducted; these functions start with the definition of short and medium term objectives of the CoP, establishing the strategies required to achieve them. The CoP members have to elaborate some documents such as a technical background document, the state of the art and practice document of the CoP domain and a calendar of activities. These documents have to be updated annually. The knowledge, experiences and results of the CoP are intended to be shared; therefore CoP members have to document them through regular reports. An important CoPs function is the analysis of relevant issues and problems of the CoP domain and to propose innovative methods of solution. Also CoPs have to identify and propose the development of innovative products and services by conducting preliminary technical and economic feasibility studies. To promote standardization across projects and departments, CoPs have to identify appropriate development platforms. As specialists, CoP members recommend actions in accordance with the direction of technology and the results obtained.

CoPs has a close interaction with the development of research projects, since the results, products and learning of a CoP are applied in their development and in turn, the knowledge generated in the project is shared in the CoP. Figure 5 depicts the functions and operation of a CoP.

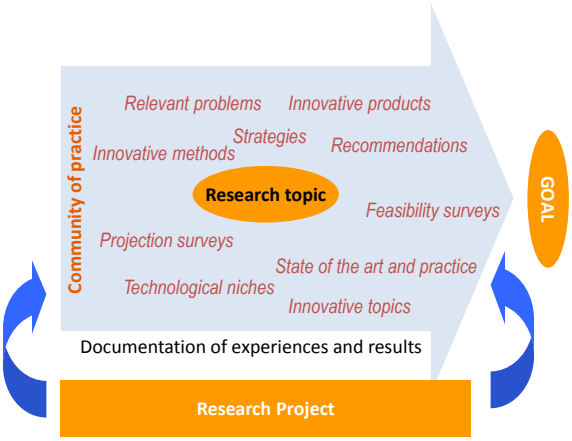


Figure 5: CoP operation and functions.

We think that for a CoP to be successful, it has to have a formal status in the organization, but members have to feel independent to work and to meet and to be creative. Therefore, we try to formalize the results they already produce before belonging to a CoP and they have autonomy to organize the work inside the CoP.

Another important point in the implantation of CoPs is resources allocated to achieve objectives.

CoPs need financial and material resources but also they need support, coaching and logistic assistance.

To support CoPs, we have implemented talks with researchers with potential to belong to a specific CoP. In these talks we try to explain the benefits of a CoP base on knowledgeable literature and successful cases. We try to fit all these theories on own experience and history, since, as we already mentioned, CoP are not new in IIE.

Thus far we have depicted the regulations and procedure for CoP, in which we tried to sensitize research staff about the impact of CoP in the projects development and in personal development. We have supported the beginning of CoPs, by helping in selection of coordinators and secretaries. Annually, we have met with each CoP to discuss about its results and resources. We have tried to monitor the functioning of the CoP.

4 WEB INFORMATION SYSTEM

In order to achieve goals, CoPs requires some means to support CoPs members collaboration, therefore an information system was developed based on Web technologies.

This system supports CoP members in basic activities, such as to register objectives, active members, technical documents and an activities calendar. Figure 6 shows a screenshot of the information system. Information about Virtual Reality CoP is shown.

The system has facilities to load current technical documents, activity reports and memorandums of meetings agreements and to keep track their lifetime. Moreover, the members can keep their own activity record and a vita which includes their production.

The system has public and private sections. In the public section non-members can ask questions to CoP members, to access open documents and to know current topics being discussed. In the private section, members can initiate blogs, monitoring activity and documents elaboration.

The activities of the projects frequently requires research work field, therefore we provide with the electronic resources for virtual meetings to use when researchers are away from office.

The principal users are: CoP coordinator, CoP secretary, CoP members and manager for the CoP process. A detailed description of the system can be found in (Argotte, 2014).



Figure 6: Communities of practice information system (In Spanish).

5 ACTIVITY INDICATORS

Since it is difficult to see whether or not a community of practice is working or it is successful just in terms of learning, we establish a list of indicators of activity and success for each particular CoP. These indicators are not supposed to be completely accomplished, but these indicators are mainly important to know if the CoP requires support or if it has to be reconfigured. Indicators allow knowing whether a CoP is working, how we have to support it and what resources are needed. Some of these indicators are:

1. CoP is composed by coordinator and secretary and at least four additional members.
2. Members understand the CoP mission and aims.
3. Members have elaborated an annual work schedule.
4. Members have elaborated a state and practice of art document and it is annually updated.
5. Members enthusiastically collaborate in the tasks assigned.
6. CoP meets twice a year at least.
7. CoP has generated ideas which have resulted in projects or projects proposals.
8. CoP members have identified and defined an innovative product.
9. CoP has a network of contacts which includes customers, suppliers, specialists, research centers.
10. CoP members have organized events where external experts are invited to share experiences.
11. CoP members have generated ideas, initiatives or proposals, which have led to the development of projects whose products, are liable to be protected by patents or copyrights, or suitable for technology transfer.

6 CONCLUSIONS AND FUTURE WORK

The IIE is conscious of the importance of having a knowledge management system and it is concerned about competitiveness and providing innovative products and services. Therefore it has implemented several mechanisms to motivate the generation and sharing of knowledge.

The communities of practice respond to these needs, motivating the knowledge sharing between colleagues. CoPs help to keep the knowledge on the cutting edge, deploy it, leverage it in operations, and spread it across the organization.

We have initiated the process of fostering CoP in the research personnel of the Enabling Technologies Division of the IIE which comprises four departments. Also, we are planning and working on expanding the model to reach the whole IIE to include all business processes.

A Web information system has been used to support CoP work, this system is being updated to include some researcher comments in order to be more supportive and user-friendly. This new functionalities includes emails alarms, chats.

In this paper, we presented the work and results at Enabling Technologies Division.

Communities of practice at IIE are the result of IIE efforts to develop and deliver innovative products and services to its customers; and it is expected CoP are consolidated in a medium term.

ACKNOWLEDGEMENTS

Authors would like to thank DTH/IIE researchers and managers for many useful discussions on the definition of the communities of practice model; also we would like to thank Guillermo Rodríguez and Miguel Pérez for their comments on previous versions of the paper.

REFERENCES

Argotte, L. 2014. *Communities of practice module*. Internal report in Spanish, Instituto de Investigaciones Eléctricas.

CFE Comisión Federal de electricidad, 2013, *The Company*. Retrieved in October 2013, http://www.cfe.gob.mx.

Davenport, T. H., Prusak, L., 2000, *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business School Press.

Hernández, Y., Arroyo-Figueroa G. and González-Castro, S., 2013. *Method for forming, update and cancel communities of practice*. Internal report in Spanish, Instituto de Investigaciones Eléctricas.

IIE, 2013. *Instituto de Investigaciones Eléctricas.* Retrieved in October 2013, http://www.iie.org.mx/

Lave, J., Wenger, E., 1991. *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press.

Wenger, E., McDermott, R., Snyder W. M., 2002. *Cultivating Communities of Practice*. Harvard Business Press, USA.

Wenger, E., and Snyder, W., 2000 'Communities of Practice: The Organizational Frontier', Harvard Business Review, v. 78, no. 1, pp.139 – 145.

UnderstandIT: A Community of Practice of Teachers for VET Education

Maria De Marsico[2], Carla Limongelli[1], Filippo Sciarrone[3], Andrea Sterbini[2] and Marco Temperini[3]
[1]*Roma Tre University, Faculty of Engineering, Via della Vasca Navale, 79, 00146, Rome, Italy*
[2]*Sapienza University of Rome, Department of Computer Science, Via Salaria 146, 00134, Rome, Italy*
[3]*Sapienza University of Rome, Department of Engineering Systems and Management, Via Ariosto 25, 00185, Rome, Italy*
{*demarsico, sterbini*}*@di.uniroma1.it*, {*limongel, sciarro*}*@dia.uniroma3.it, marte@dis.uniroma1.it*

Keywords: Communities of Practice, Social Networks.

Abstract: Social Networks are among the most popular achievements of Web 2.0. As a matter of fact, the Internet is full of social communities like Facebook or Twitter or others, which are used not only by teenagers but also by older users to share experiences, information and opinions. The expression Social Network allows for a wide interpretation, coming from the research in social and behavioral sciences. It also lends itself to encompass a kind of more thematically vertical communities, namely Communities of Practice, where groups of people interested to learn a particular practice are supported in achieving this goal thanks to the professional relationships with expert peers. The community network evolves with time, as novice users become experts in turn, and new communication protocols arise. We start from the consideration that most high school students nowadays use Web 2.0 instruments and tools for chatting, streaming or posting messages. An interesting challenge is the possibility for teachers the use of the same instruments and tools to enhance their daily work. To this aim, we present a first study of a Community of Practice of teachers, involved in Vocational Education and Training (VET), aiming at learning the use of Web 2.0 tools and applications to support the learning process of their pupils. In practice, we propose to use Web 2.0 to learn to teach through Web 2.0. As a platform we used the web application ELGG, that is an open source social network engine thanks to which we built a social environment where a sample of VET teachers participated to several social activities. A first evaluation of effectiveness of such an approach is encouraging.

1 INTRODUCTION

The expression Social Network was born in the field of social and behavioral sciences and soon acquired a weight of its own. A lot of research has been spent to analyze these social structures made up by a set of individuals or organizations and their (dyadic) ties (Wasserman and Faust, 1994). Thanks to technological progress and, most of all, to the spread of the Internet and to the rise of Web 2.0 protocols and frameworks, the same expression is nowadays mostly referred to virtual places for sharing information, opinions, data and whatever among connected peers of a community. In practice a Social Network (SN) is a set of people or groups of people with similar pattern of contacts or interactions such as a friendship, co-working, or information exchange (Radicchi et al., 2004). In this scenario, individuals register into virtual communities for different purposes: to know other people, to share daily experience with *friends* in a sort of shared virtual diary, but also to exchange professional knowledge, to learn about particular topics and so on. Among the most populated SNs, we can mention *facebook*[1] for contacts, texts or messages sharing, *YouTube*[2] for video sharing, *linkedIn*[3] for job posting and sharing and their number is continuously growing. *Communities of Practice* (CoPs) can be considered as an especially vertical kind of SNs where members enroll to reach some goals related to very specific topics in particular domains, typically concerning their work or even a kind of hobby (Wenger, 2010). CoPs are groups of people who share an interest or a passion for something they do, and aim at learning how to do it better by interacting regularly. They are not merely clubs of friends or networks of connections between people, but rather own an identity defined by a shared domain of interest (from the Wenger web site [4]). Here we present a par-

[1]www.facebook.com
[2]www.youtube.com
[3]www.linkedin.com
[4]http://wenger-trayner.com/theory/

ticular CoP, called *UnderstandIT*[5] from the European project inspiring it, having Italian *Vocational Education and Training* (VET) teachers as members. Their common goal is the learning of the best practices to use some Web 2.0 instruments and tools, such as forum, chat and wiki, in their daily teaching activities. This project arose from the need to bridge the gap between the use of Web 2.0 by teachers, compared to that by pupils, and to make digital immigrants (the former) and digital natives (the latter) encounter in a common ground (Prensky, 2001). In fact, the underlying hypothesis is that the use of these tools has great potentialities to make teaching more effective. In the original European Project, teachers of different European countries (Lithuanian, Italians, Portuguese, German, Danish, and English) used a CoP to reach the same professional goals. The UnderstandIT CoP was designed following the Wenger's schema, based on the concepts of *domain*, *community* and *practice* (Wenger, 1998). The domain of shared competence addressed here is the teaching activity for VET education; the community members are VET teachers while the practice is the use of the Web 2.0 instruments and tools in the teaching activities. In addition, peer-mentoring as well as peer-coaching are adopted, since they are widely recognized as winning strategies, most of all when learners are teachers in their everyday life. However, coaching is an art, that implies carefully tuned presence/absence of the coach, to spur coachee's abilities regarding critical planning, and an accurate evaluation of obstacles and related overcoming activities. We used the open source ELGG social network engine[6] as a platform, since it is one the most used frameworks delivering the building blocks that enable companies, schools, universities and associations to create their own fully-featured social networks and applications. This engine runs as a web application, providing a social environment with a wide range of Web 2.0 services such as forum, chat, wiki and so on: members are free to create wiki activities and to participate to all the social activities put on line by the platform. The research question of our work is to test if such a *reflexive* approach to the learning of Web 2.0 really helps the teachers of the UnderstandIT community to learn and to teach. The powerful message carried by the learner centric model can be enhanced by enriching the role of the teacher to include greater individual support and guidance for the learners (Quintana et al., 2001). The remainder of the article is structured as follows. Section 2 draws some important related work; Section 3 shows the system, i.e., the social network engine which the CoP is based on. In Section 4 the social network along with its evolution is shown while Section 5 presents the experimental evaluation of the overall system. Finally in Section 6 some conclusions are drawn.

[5]http://understandit.di.uniroma1.it/

[6]http://elgg.org

2 LITERATURE REVIEW AND RELATED WORK

Even if it is not possible to determine the exact birth of the modern interpretation of CoPs (see for example (Bourdieu, 1977) and (Vygotsky, 1978)), it is indubitable that its success has been significantly supported by the works by Lave and Wenger (Lave and Wenger, 1991; Lave, 1991). The main stream of innovation has been related to riding out the idea that work practice must be conservative and stable in time and learning must be taken separate from it and be focused on the most static aspects. One of the factors supporting this has been the concrete understanding that the ways people actually work are usually different from what described in manuals, training contents, and job *theoretical* descriptions. Differences are related to both complementary as well as completely alternative strategies and rules. Nevertheless, it has been noticed that organizations tend to rely on the *static* form of training in their attempts to understand, improve, and transmit work practice. On one hand, this kind of behavior was generated by a society structure recognizing a special value to *abstract knowledge*: the practice details are considered as contingent, and therefore less essential, or at least easily derived after the relevant abstractions. It is to say that the above conventional theory of learning and most of all of training is far from the ancient institution of the artisan workshop, with their primary role in transmitting and preserving arts and crafts. However, the industrial era took this prevalence of abstract knowledge over actual practice and to separate learning from working, as well as learners from workers. By contrast, the new theories returned to the *ancient wisdom*, suggesting that appropriate practice is central to work understanding, and that abstractions detached from practice might even distort or obscure less trivial details of practice. Further, they assigned a special role to social exchange of experience and information. Transferring these considerations to learning was quite immediate, also due to the theories of Vygotsky (Vygotsky, 1978) and of the other researchers in the line of socio-constructivism. Lave and Wenger built on a practice-based theory of learning, which has its core concept in the *Legitimate Peripheral Participation* in CoPs (Lave and Wenger, 1991). The novices enter the community from the pe-

riphery, yet over time move towards full participation by acquiring experience and reputation. However, the community itself also moves through stages of development which are characterized by different levels of interaction among the members, and also by different kinds of activities. This concept provided one of the most versatile accounts of the constructive view of learning, even if it is not actually a method of education. Its focus is on learning, rather than on teaching (Wenger, 2010). It is interesting to consider that a topic formerly focused on organizational learning, at present deserves an increasingly significant role in any learning field. Moreover, firstly the spreading of the Internet, with early forms of forums, and secondly of modern SNs, supported by Web 2.0 technologies, have further supported the possibility and spurred the inclination to exchange on virtual communities not only personal data but also working experience and learning difficulties. With the creation of dedicated software frameworks, CoPs have become one of the most appreciated forms of social learning (Blackmore, 2010). In recent projects CoPs have also been used to implement career education networks, aiming at supporting an aware transition from school to work and/or further education, requiring a lot of information from a wide range of sources. In this new perspective, learning is both the main activity and the core topic which is supported by the CoP strategies. Buysse, Sparkman and Wesley (Buysse et al., 2003) underline *the growing need to integrate educational research and practice* to connect what we know with what we do. However, they underline that shared lists of recommended practices often fail to promote the personal responsibility and exploration ability of all involved stakeholders, including not only educational researchers and teachers but also parents and students. Every member of the educational community should be encouraged and motivated to analyze and possibly criticize, in a constructive way, each other experiences and teachers should be engaged in research activities as well. We consider CoPs a very promising environment to reach this goal, to joining-research and practice, among which they mention action research (Calhoun, 1994), or professional development schools. The trick is breaking *the linear relationship through which information is handed down from those who discover professional knowledge to those who provide and receive educational services* (Buysse et al., 2003). The two central elements of the CoP approach are: i) situated learning (Hummel, 1993) and ii) community reflection on practice. Knowledge is acquired from and applied to everyday settings, and is discussed with peers and experts that made up a social system (Wenger, 1998). Other approaches concern: i) the personalized learning where the added value is given by a didactic capable to tailor its patterns to the learners' specific needs, such as in a formal education (see for example (Limongelli et al., 2011a; Limongelli et al., 2008a; Limongelli et al., 2011b; Sterbini and Temperini, 2010) for an insight into this area) ;ii) the ontological approach where teachers are helped by ontology-base systems in order to search for suitable didactic material in the Internet (see (Limongelli et al., 2010; Limongelli et al., 2012; Limongelli et al., 2013b; Limongelli et al., 2013a)). Moreover, in a perspective of technology enhanced learning, there is some research work (De Marsico et al., 2011; De Marsico et al., 2012) aiming to integrate more traditional individualized e-learning and social-collaborative e-learning (Sterbini and Temperini, 2009; Limongelli et al., 2008b). The project closest to our work is the *SEDA* project, that is a network of people engaged in staff and educational development. In (Nixon and Brown, 2013) a CoP for educational developers has been proposed, highlighting the needs for teachers of sharing their experiences. In our work we focus our attention on a specific objective that is the involvement of teachers in the use of Web 2.0 tools as vehicles for captivating learners.

3 THE SYSTEM

The design of the system started from the requirement analysis, based on an input coming from a sample of VET teachers to identify a list of useful features to deploy:

- Groups common goals sharing and shared editing tools;
- Friends, groups, stream of interactions and event calendar and RSS feed import;
- Antispam protection;
- FAQ managing;
- Activities sharing among friends;
- Embedding external contents;
- Tracking changes/hot topics;
- Synchronous (chat) and asynchronous (blog, page, comment, tweet, forum) communications.

Subsequently we analyzed several open source environments and selected the ELGG framework. ELGG is a *LAMP* (Linux, Apache, MYSQL, Php) web application that provides a basic set of services for a social network system, allowing for a complete interface configuration and extension of the functionalities (by adding plug-ins that are either newly implemented or

selected among the many already available from the ELGG development community). Finally, the system has been deployed and customized to satisfy the above requests. In SNs members share and exchange knowledge by adding a new bookmark, posting a link to a particular video, publishing new pages and so on and people might feel the need to know those members of the community that are interested in sharing that knowledge. So, among the several features offered by a social networking platform, of special relevance is the setting of access rights associated to each contributed resource. In particular, the author of a contributed resource can specify a *group* (a *class of users*) to whom release either *read access* or *read and write access* rights on the resource. The *group* is a subset of the CoP members; there are several possible group specifications available (listed from a more public to a more private nature): *Everybody*, *All registered users*, *A Group*, *All your friends*, A *subset of your friends* and *Private*. Depending on the write access rights of a post, the contributed contents can be edited asynchronously by different users. Each post in the system can be tagged by the author, i.e., it can be associated with one or more keywords. This allows for searching through the system resources by both *full text* and *tag-based* selection. Another characteristic of the system is that the UnderstandIT CoP has been configured in such a way to allow for discussions in several language-specific communities, as mentioned in Section 1, with a common discussion space in English. This feature has been implemented through the concept of *Group* managed in ELGG, where a Group is a subset of all the CoP members that is owned by a member and can be moderated. Each group represents a (sub-)community whose members interact through:

- Group Blog posts, Group Bookmarks and files and Group pages (in particular the editing of subpage trees was made easier through an appropriate plugin);
- Possibility to embed videos/podcasts and RSS feeds and to upload resources through the *File Tool* feature (with the organization of files supported by the File trees plugin);
- Possibility to embed external content by reference through the Media Tool;
- Possibility to add (asynchronously) comments on posted resources (files, media, bookmarks and blogs);
- Possibility to share interesting resources through the Bookmarks Tool;
- Possibility to monitor (import) RSS feeds through the Feed Tool;
- Management of Forums and Chats.

Moreover, in order to increase the support to collaboration, a synchronous shared editing plug-in has been installed. It is based on the *Etherpad* open source project [7]. Through this plug-in the Group members can share the editing process on the same document, i.e., members can cooperate (also by using a small chat) through:

1. Normal (wiki-like) pages, with the support to comments and revisions;
2. Shared (etherpad) pages, which allow for: i) synchronous editing as *Google docs*, where each editor's input is differently highlighted; ii) replay to all changes/revisions; iii) export/import to/from documents in various formats (rtf/doc/odt/html/pdf).

Users of the CoP are kept updated about the CoP activities either through automatic e-mail notifications or by means of the CoP web interface. Both the former and the latter collect information from many relevant areas of the network, and offer appropriate notifications, such as in the following:

- E-mail, RSS feeds, status tweets;
- Last changes in page history;
- Group activities;
- Last changes in the user's *dashboard*;
- Last changes in the main CoP's page and in each Group main page;
- Contextual tag cloud providing information about the terms used in the CoP and their location.

To make spammers' life harder an ELGG plug-in has been selected and installed and members can report occurrences of spamming to the administrator, who in turn can remove the spammer and add username and e-mail to the global `stopforumspam.com` database. To externally disseminate the CoP discussions towards Twitter, a plug-in has been installed to broadcast twits generated by the CoP users.

Since the original UnderstandIT CoP was born in the context of the European project with the same name, its interface was to the languages of the UnderstandIT project partners (Lithuanian, Italian, Portuguese, German, Danish, and English). To allow for cooperation among different language groups, an automatic Google Translation-based tool has been installed in all pages. Partners have found the Google-translator tool sufficient to understand the knowledge posted by other language people. This was an interesting experiment for international courses, and for inclusion of contents in foreign languages.

[7] http://etherpad.org/

4 THE CoP ACTIVITY

In this section we show some main activities of the CoP members. We describe participants' interactions about their collaborative activities by means of the social Web 2.0 communication tools described in the previous section, i.e., forum, blogs, faqs, and so on. In particular, the UnderstandIT CoP was used as a container for an online course deployed adopting peer-to-peer coaching and mentoring strategies. For the reader's sake, from now on we will use UnderstandIT as a name for the course too. This course was attended by a sample of 64 teachers, coming from different didactic experiences, as shown in Fig. 1. Intentionally we randomly selected the sample of teachers from different educational institutions in order to obtain a significant set of teachers.

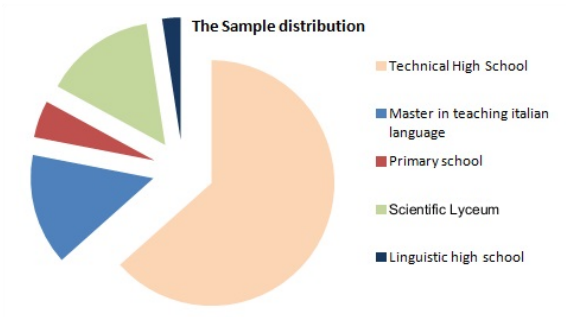


Figure 1: The sample distribution of the didactic work experiences of the CoP members.

The overall learning outcome was: *I can choose relevant resources using the CoP approach for any planned learning activity including ICT-tools where these are the most appropriate and I am able to change my teaching style in order to use a more coaching oriented approach.* In the following we show the detail of learning outcomes of the course that have been proposed in English and Italian.

Detailed Learning Outcomes

1. Information sources: I can help my students to find and evaluate relevant information including pictures, maps, sound and video from ICT sources as well as traditional sources. (e.g. RSS-Feeds, Wikipedia);
2. Information storage: I can help my students to store information in digital form such as blogs, wikis, podcasts or portfolios (so it is accessible at home, on the move and to external experts/advisors);
3. Communication: I can help my students to communicate synchronously and asynchronously through text, voice and video;
4. Documentation: I can author online spaces to document learning events and facilitate student use of this material afterwards (e-portfolio, screen casts, class blog/wiki);
5. Safe use: I can help my students to use online applications safely (e.g. online identity, privacy, trustworthiness);
6. Group work: I can facilitate online group working, using methods such as complex instruction;
7. Learning: I can identify the learning outcomes by using specific ICT tools in a specific learning activity;
8. Assessment: I can identify how the learning outcomes of ICT related activities will contribute to student assessment and use appropriate rubric tools to record this;
9. Planning: I have made a plan for my continuing professional development in the use of ICT in my teaching;
10. Documentation: I have documented the use of an interactive internet tool so that others can benefit of this;
11. Sharing: I have tried at least one way of sharing my ICT experience with either internal or external colleagues;
12. Local context: I have identified possibilities/limits of the use of interactive Internet tools in my institution;
13. Portfolio: I have started or added to an existing e-portfolio;
14. Evaluation: I have evaluated the outcomes of my amended teaching session and documented this in my e-portfolio;
15. Coaching: I have embarked on peer coaching with one of my colleagues;
16. Coaching: I have the skills of listening and questioning to be able to coach other people.

The main achievements of such a course were basically in the deeper appreciation of the amount of work required from teachers in order to timely deal with the assigned tasks, and in the knowledge about which parts of the course need localization and/or translation, given that this is feasible (e.g., external material in pdf is heavy to completely translate). The course included several tasks such as:

- Practice with Web 2.0 tools for doing and analyzing one's own teaching;
- Analysis of existing CoPs;
- SNs and blogs to coach sessions with the course tutor and with peers;
- Final composition of a plan to experiment one's learning outcomes in her present teaching activity.

The visits of the CoP members at the web site, over the entire duration of the course, were about 2,000, distributed as shown in Fig. 3, i.e., in average 31.25 visits for member. In Fig. 4 the tag clouds concerning the Blog and the Bookmark activities are shown. Finally, in Fig. 5 we show the overall CoP activity in terms of a standard metrics like the total number of posts in the forums, the total number of faqs and so on. All the above activities show that the members

have been engaged enough in the CoP life. In particular for the blog activity we see that there have been posted 90 discussions and 700 posts as shown in Fig. 5. Fig. 2 shows a view of the system interface and an example of a group interacting through group blog.



Figure 2: At the bottom an example of interaction among peers, at the top the home page of the course.

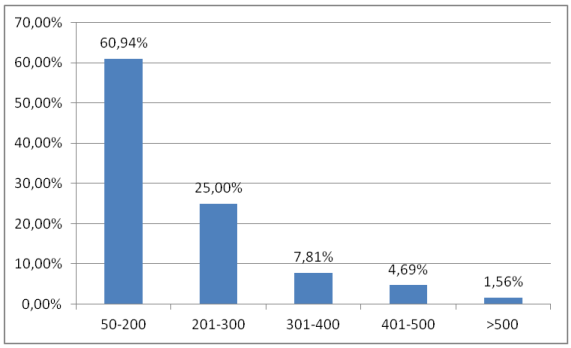


Figure 3: The total number of visits.



Figure 4: The tag cloud of the bookmark activity.

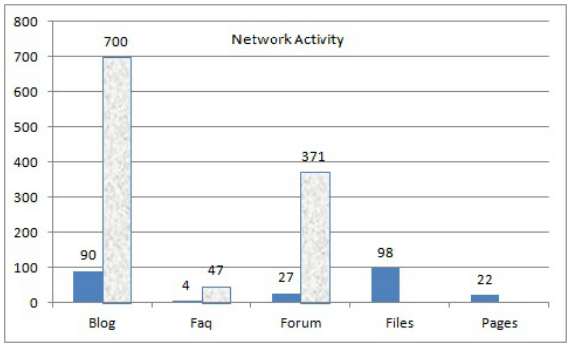


Figure 5: The overall CoP activity. Blue and grey bars for Blog, Forum and Faq represent their total number while their grey bars represent the total number of answers to them. Files and Pages bars represent their total number.

5 A FIRST EVALUATION OF THE CoP

In this Section we present a first descriptive evaluation of the CoP with the aim to measure the degree of satisfaction of the members with respect to the community activities and reasoning on the effectiveness of the CoP. Evaluating a CoP is not a trivial task: it is easy to get overwhelmed with the variety of evaluation approaches and metrics available. Once an online community of practice becomes active, leaders need to develop an evaluation plan to understand and document success in their community. In many contexts, evaluation approaches, which measure communities activities and performance, are also referred to as *metric analysis*. Here we show the evaluation of the CoP performance by means of two key activities submitted to the CoP members, represented by a pre- and a post-test questionnaire.

5.1 The Pre-test Questionnaire

Through the pre-test questionnaire we measured some general variables like the members experience on the use of Web 2.0 tools. It was based on 16 questions concerning mainly the learning goals, the starting ICT and Web 2.0 skills and the work experience. Interesting questions were:

1. Have you ever participated to online courses to date?
2. Have you ever participated in Communities of Practice online?
3. Which of these tools (forum, chat, blog, rss,...) do you use or have used at least once?

More than 70% of the members had already participated in an online course and in the same way the majority of the sample, i.e., the 71%, had not experience in CoPs. On the other hand, Fig. 6 shows that the majority of the members had experience in the use of some Web 2.0 instruments like forum, chat and blog.

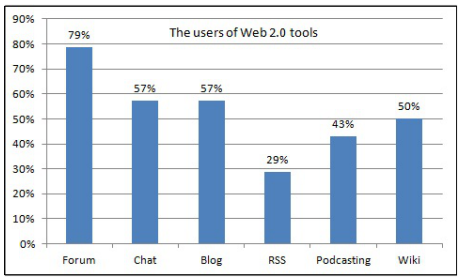


Figure 6: The answers to question 3: the Web 2.0 skills.

According to the above considerations, taking into account what shown in Section 4 as well, we can conclude that the sample of the CoP members was very articulated with an average skill on Web 2.0 tools.

5.2 The Post-test Questionnaire

Here we present the results of a questionnaire submitted to the CoP members after having taken the course. It was composed of 21 questions concerning different aspects of the didactic experience. We report two important questions:

1. Did the course meet your expectations?
2. Do you feel yourself prepared to act as a coach to your colleagues in the use of Web 2.0 tools?

We used a 5-points likert scale to measure the answers. From Fig. 7 we see that the 62% of teachers has been satisfied by the course: it met their expectations. This is a very important key for the success of any course. Finally, Fig. 8 is a first proof of the effectiveness of the course. We discussed with the teachers about an aspect which is worth mentioning. The *reflexive* use of Web 2.0 to teach Web 2.0 technologies was found particularly engaging, so that this can be considered as a point in favour of adopting this approach on a larger scale. The added value is that using the same tools that the student is learning to exploit reinforces the final achievements and is much more engaging, along the line of *learning by doing*.

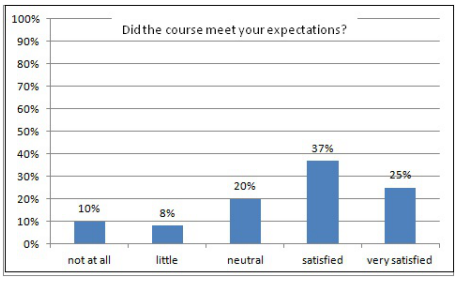


Figure 7: The answers to question 1: the expectations.

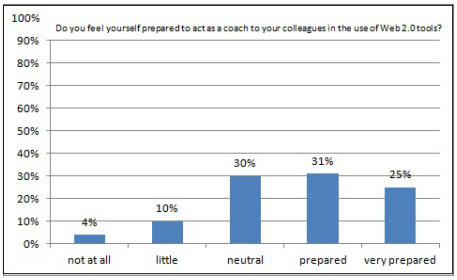


Figure 8: The answers to question 2.

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented a CoP of teachers, called UnderstandIt, to learn the Web 2.0 instruments and tools. We implemented the CoP by means of the social engine ELGG and the community ended its learning process few time ago. Here we show the first results of such an experiment, showing both the network activity and an evaluation process, performed by means of the submission to the CoP members of two questionnaires, the pre-test and the post-test questionnaires. The first results are encouraging: the majority of the members feel prepared to coach other colleagues to teach Web 2.0 instruments and tools in their didactic activity and feel the course met their expectations. As a future work we plan to do a deeper analysis of the log data together with taking into account all the statistical correlations between the pre

and the post activity questionnaire in order to make other measures of effectiveness, test-based as well. Also we plan to repeat the experiment in a broader context.

REFERENCES

Blackmore, C. (2010). *Social Learning Systems and Communities of Practice*. The Open University.

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge: Cambridge University Press.

Buysse, V., Sparkman, K., and Wesley, P. (2003). Communities of practice: Connecting what we know with what we do. *Exceptional Children*, 69(3):263–278.

Calhoun, E. (1994). *How to use action research in the self-renewing school*. Alexandria, VA: Association for Supervision and Curriculum Development.

De Marsico, M., Sterbini, A., and Temperini, M. (2011). The definition of a tunneling strategy between adaptive learning and reputation-based group activities. In *Proc. 11th IEEE Int. Conf. on Advanced Learning Technologies, ICALT*, pages 498–500.

De Marsico, M., Sterbini, A., and Temperini, M. (2012). A strategy to join adaptive and reputation-based social-collaborative e-learning, through the zone of proximal development. *International Journal of Distance Education Technologies*, 19:2:105–121.

Hummel, H. (1993). Distance education and situated learning: Paradox or partnership? *Educational Technology*, 33(12):11–22.

Lave, J. (1991). Situating learning in communities of practice. In Resnick, L., Levine, J., and Teasley, S., editors, *Perspectives on Socially Shared Cognition*, chapter 4, pages 63–82. American Psychological Association, Washington, DC.

Lave, J. and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.

Limongelli, C., Lombardi, M., Marani, A., and Sciarrone, F. (2013a). A teacher model to speed up the process of building courses. In *Human-Computer Interaction. Applications and Services - 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II*, pages 434–443.

Limongelli, C., Lombardi, M., Marani, A., and Sciarrone, F. (2013b). A teaching-style based social network for didactic building and sharing. In *Proceedings of AIED 2013 - 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013.*, pages 774–777.

Limongelli, C., Mosiello, G., Panzieri, S., and Sciarrone, F. (2012). Virtual industrial training: Joining innovative interfaces with plant modeling. In *Proc. of ITHET 2012*, pages 1–6. IEEE.

Limongelli, C., Sciarrone, F., Starace, P., and Temperini, M. (2010). An ontology-driven olap system to help teachers in the analysis of web learning object repositories. *Information System Management*, 27(3):198–206.

Limongelli, C., Sciarrone, F., Temperini, M., and Vaste, G. (2008a). Lecomps5: A web-based learning system for course personalization and adaptation. In *Proceedings of IADIS 2008, Amsterdam, The Netherlands, July 22-25, 2008. Proceedings*, pages 325–332.

Limongelli, C., Sciarrone, F., Temperini, M., and Vaste, G. (2011a). The lecomps5 framework for personalized web-based learning: A teacher's satisfaction perspective. *Computers in Human Behavior*, 27(4):1310–1320.

Limongelli, C., Sciarrone, F., and Vaste, G. (2008b). Ls-plan: An effective combination of dynamic courseware generation and learning styles in web-based education. In *Proceedings of AH'08: Adaptive Hypermedia and Adaptive Web-based Systems*, volume 5149 LNCS, pages 133–142.

Limongelli, C., Sciarrone, F., and Vaste, G. (2011b). Personalized e-learning in moodle: The moodle-ls system. *Journal of E-Learning and Knowledge Society*, 7(1):49–58.

Nixon, S. and Brown, S. (2013). A community of practice in action: Seda as a learning community for educational developers in higher education. *Innovations in Education and Teaching International*, 50(4):357–365.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon*, 9(5):1–6.

Quintana, C., Carra, A., Krajcik, J., and Elliot, S. (2001). Learner-centred design: reflections and new directions. In Carroll, editor, *Human-computer interaction in the new millennium.*, pages 605–626. ACM Press.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658.

Sterbini, A. and Temperini, M. (2009). Collaborative projects and self evaluation within a social reputation-based exercise-sharing system. In *Proc. Int. Conf. on Web Intelligence and Intelligent Agent Technology - WI-IAT Workshops*, volume 3, pages 243–246.

Sterbini, A. and Temperini, M. (2010). Selection and sequencing constraints for personalized courses. In *Proceedings - Frontiers in Education Conference, FIE.*

Vygotsky, L. (1978). *Mind in Society: Development of Higher Psychological Processes*. Cambridge MA: Harvard University Press.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge Un. Press.

Wenger, E. (1998). Communities of practice: Learning as a social system. *Systems thinker*, 9(5):2–3.

Wenger, E. (2010). Communities of practice and social learning systems: the career of a concept. In Blackmore, C., editor, *Social Learning Systems and Communities of Practice*, chapter 11, pages 179–198. The Open University.

Change Management in Service Oriented Virtual Organizations
A Structural Framework to Identify the Impacts and Triggers of Changes

Waeal J. Obidallah[1,2] and Bijan Raahemi[1]
[1]*Faculty of Graduate and Post-Doc Studies and Telfer School of Management, University of Ottawa, Ottawa, Canada*
[2]*Collage of Computer and Information Sciences, Imam Mohammed ibn Saud University, Riyadh, Saudi Arabia*
{wobidall, braahemi}@uottawa.ca

Keywords: Virtual Organizations, Service Oriented Architecture, Change Management, Change Classification, Triggers of Changes, Levels of Changes.

Abstract: Business processes and services of Service Oriented Virtual Organizations are subject to change to meet the internal and external requirements of the competitive, complex and rapidly changing collaborative environment they operate in. Efficient and practical change management solutions are needed to enable the partners to gain insight on the various triggers of change, and their impacts on the Virtual Organization's operations and collaboration. This paper presents a structural change management framework to facilitate the process of change by allowing the participating partners in a Virtual Organization to identify and understand the levels and the triggers of changes in their environment. Our proposed solution consists of the structural framework including two layers, namely, the levels of changes, and triggers of changes. We present an example scenario to demonstrate how employing the proposed framework to identify, recognize and categorize various changes in the Service Oriented Virtual Organization improves the flexibility, and agility of the change management in the organizations.

1 INTRODUCTION

The evolution of web technologies and services has transformed the internet from a mere repository of information and data into a useful and an inevitable platform for service provisioning and sharing. It enables businesses from different locations to share information, and resources in the form of Collaborative Networked Organizations (CNOs) where they share risk, skills, resources and core competencies to add more competitive advantages to their organizations (Akram et al., 2010). The result of this is the emergence of the virtual organizations (VO), which is a dynamic, temporal consortium of autonomous legally independent organizations that collaborate with each other to achieve certain objectives and meet business needs (L. M. Camarinha-Matos et al., 2008). The idea of aligning VOs with service oriented architecture (SOA) is potentially one of the best ways to implement, and manage dynamic business processes which led to the concept of a service oriented virtual organization (SOVO) (Danesh et al., 2011).

Since today's business environments are very competitive, complex, and rapidly changing; companies and different organizations are increasingly restructuring, changing, and improving their internal business processes, services, information systems, and partnerships with other companies to leverage, and add more value to their businesses (Camarinha-Matos et al., 2008). The VO participating organizations need to evolve to meet the market and customer demands by changing their own business processes and services. Thus, the VO partners need to understand the levels, impacts, triggers and the consequences of changing their business processes and services to meet the market and customers' demands. Consequently, change management processes and procedures are needed to facilitate and manage the process of change in a collaborative way between different partners in the SOVO environment (Obidallah et al., 2013).

In this paper, we present a structural change management framework by identifying the levels and triggers of changes which improve the flexibility and agility in changing shared-services and business processes between different partners in the SOVO, enabling its partners to understand the triggers, categories, and impact of changes in their collaborative environment.

The rest of the paper is organized as follows: In Section 2, we introduce our research motivation. In

Section 3, we describe changes in the Service Oriented Virtual Organizations (SOVO) by referring to the concept of virtual organization (VO), service oriented architecture (SOA) and change management. This is followed by our proposed structural change management framework in Section 4. Example scenarios of changes in the SOVO are described in Section 5. Finally, we conclude the paper in Section 6.

2 MOTIVATION

VOs are dynamic and continuously changing because of the dynamic environments in which they operate (Liu and Bouguettaya, 2007). The idea of combining the SOA with the VO presents many advantages in today's business environments, but it also introduces new research challenges. Industry analysts state that one of the leading causes of downtime and missed Service Level Agreement (SLA) is unmanaged change. Furthermore, the Gartner Group stated that enterprises should invest more to improve change management processes, problem management processes and automation tools to address 80 percent of unplanned downtime. This will help organizations to reduce downtime caused by application failures and operator errors (Dumitraş et al., 2007). The various changes in the SOVO environment are expected to be organized and managed in a way that allows participated partners to evolve and change their own business processes and services accordingly. Therefore, a framework that identifies and clarifies the impact and triggers of change, in addition to the processes and procedures for handling all levels of changes plays a central role in the successful deployment, operation and evolution of the SOVO.

Consider a VOWireless where a Cell Phone provider and a Wireless Network provider collaborate with each other to provide a final product which is a cell phone with activated wireless plan delivered to the customer. The Cell Phone service provider and the Wireless Network service provider share their business process and service together to form a Virtual Organization, called VOWireless. Figure 1 illustrates the value network of VOWireless. The value network focuses on a group of entities who work and collaborate together to achieve certain business goals. It models the business value between different organizations to form a complex chain that includes multiple services providers who interact and communicate with each other to increase business value exchange (Ul Haq and Schikuta, 2010).VOWireless can expect changes to its business processes and services from different sources, where changes could be triggered for different reasons. For instance, the performance of the VO might be below the acceptable threshold, or a new business strategy might be defined. If the performance level of cell phone service provider does not meet the SLA, then VOWireless needs to react to this trigger of change, and understand how this change affects the VO operations and collaboration. The change can impact the VO in different levels depending on the severity of the change. Each change within VOWireless will affect the business process and services shared in the VO in different aspects and levels, as well.

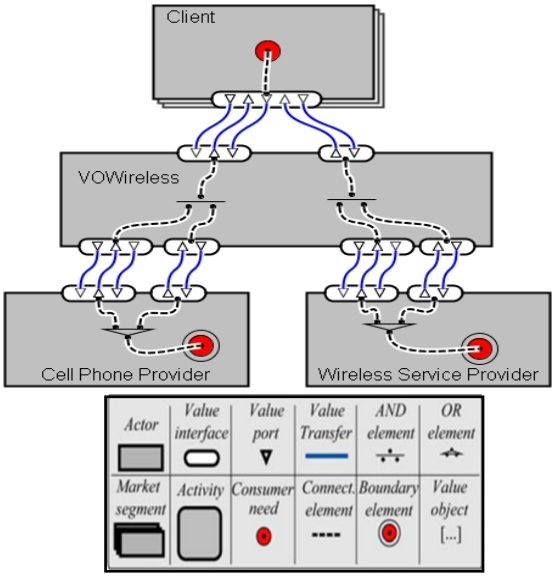


Figure 1: The VOWireless value network.

Since VOs faces and initiates different kinds of changes in order to meet the markets' and customers' demands, a VO such as VOWireless should employ a change management mechanism which identifies the levels, impact and triggers of changes, and facilitates the process of changing business processes and services by engaging all partners in the change process in a collaborative and efficient way. In this paper, we are focusing on identifying the levels, impact and triggers of changes in the SOVO environment.

3 BACKGROUND

3.1 Virtual Organization

Collaborative Networked Organization (CNO) has emerged as a result of greater awareness on the potential benefits and advantages offered by networking, partnerships and collaborations between different organizations to meet requirements (Camarinha-Matos et al. 2008). A Virtual Organization (VO) is a dynamic, temporal consortium of autonomous legally independent organizations that collaborate with each other in a way to attend a business opportunity, and meet specific requirements. VOs are very important manifestation within CNOs (Drissen-silva and Rabelo, 2008). The VO participating partners share risks, costs, and benefits, coordinate the sharing of skills, resources and competences to achieve their operations and use computer networks to support their interactions (Drissen-silva and Rabelo, 2008; Camarinha-Matos et al., 2008). The VO looks and performs its tasks and activities as if it was a single organization even though each VO participating organization is independent by having its own objectives, internal processes and business culture. This fact makes it challenging to manage the VO like a single organization (Jansson et al., 2008).

VOs have four phases in their lifecycle as show in Figure 2 (Drissen-silva and Rabelo, 2008). The VO management is “the organization, allocation and co-ordination of resources, their activities as well as their inter-organizational dependencies to achieve the objectives of the VO within the required time, costs and quality frame” (Negretto et al., 2008). Its challenges can be categorized into distribution, dependencies and coordination challenges, change management challenges, performance management challenges, knowledge and information management challenges, risk management challenges, and decision making challenges (Karvonen et al., 2004; Jansson et al., 2008).

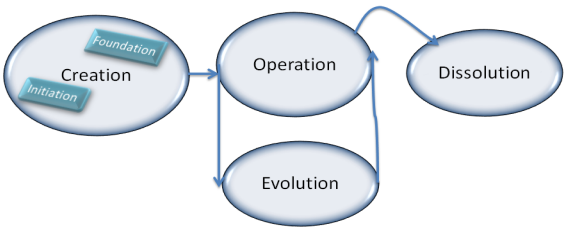


Figure 2: The VO’s Life-Cycle.

3.2 Service Oriented Architecture

Service Oriented Architecture (SOA) is an architectural style where systems consist of service users and service providers (Bianco et al. 2007). The SOA provides interoperability between different services through the use of different standards; loose coupling that minimizes dependencies and just requires the awareness between services, and abstraction which hides the logic of service from the outside world using technologies such as XML, reusability which is the ability to reuse a certain services in different business processes, autonomy which allows services to have control over the logic they encapsulate which is governed by policy and, finally, discoverability where metadata are used to discover a service (Erl, 2005). Building business applications using the SOA increases the agility and flexibility of business processes, the interoperability between different services and their reusability The SOA is the proper way to create dynamic and collaborative applications that are built for change because it is able to adopt with the dynamic business environment. It also reduces the complexity and rigidity of developing IT or business applications compared to other ways such as client-server where changes may require rebuilding the whole application again (Holley and Arsanjani, 2010).

3.3 Change Management

Our research is primarily focused on the technical aspect of change management. Change Management (CM) from an IT perspective is "a process whereby changes to a service are formally introduced and approved before deployment into the next testing or production state (Holley and Arsanjani, 2010)." It focuses on changes to a service through its whole lifecycle and ensures the prompt and efficient handling of all changes through the use of standardized methods and procedures based on industry best practices (Holley and Arsanjani, 2010). The goal of CM is to support the process of change, enable the traceability of changes, identify and minimize the impact of change and gradually improve the day to day operations (OGC, 2007). Organizations believe that virtual business models, such as the SOVO, include innovations with the potential to respond to complex and dynamic changing business environments, as well as provide benefits and various opportunities that are not found in physical organizations (Abuelmaatti and Rezgui, 2008).

Business processes and services in SOVOs are

subject to changes in order to evolve and improve its operations to meet markets and customers' requirements. The process of managing changes becomes more complicated and complex as the number of services or service providers increases dramatically (Akram et al., 2010). Managing changes between different partners manually is not practical especially with the growing number of service providers and the improvements in Information and Communications Technologies (ICTs). A more practical and efficient way is to automate the process of change between different partners allowing a quicker and user transparent way of CM. This will allow a quicker response time for changes and less efforts from the VO manager to track, maintain and coordinate for change (Akram, 2005). During the operation phase of VOs different kinds of unexpected events take place for different reasons which may cause a problem to maintain the VO's operations and achieve its objectives (Drissen-Silva and Rabelo, 2009). The unexpected events may lead to changes in the VO that require decision making, problem solving, discussion making and collaboration between independent and autonomous partners (Drissen-silva and Rabelo, 2008). A specific change to a service or business process may force a series of changes due to the dependency between services and business processes (Wang et al., 2010). The dependency between different business processes and services, in addition to the complexity and the characteristics of a SOVO makes the process of change more challenging and complicated. Therefore, change management is an important research subject which needs to be investigated to enable the VO partners to adapt to the dynamic and changing environment. The CM processes will facilitate change while also reducing user effort which eventually leads to faster VO change adoption.

4 PROPSED STRUCTURAL FRAMEWORK FOR SOVO CHANGE MANAGEMENT

Our proposed change management structural framework consists of two layers which are: (a) the levels of changes; and (b) the triggers of changes in the SOVO. These two layers of the structural framework are described in more details in the following sub-sections.

4.1 The Levels of Changes

Changes in the SOVO are triggered for different reasons and in different levels during the operation phase of the VO. These changes could start from different sources to meet the VO goals and customer expectations. The fact that the three modelled layers of the SOVO; the value network layer, the collaborative process layer and the service providers' layer, are dependent to, related, and connected with each other, makes changes in the SOVO more complex. This complexity is due to the fact that a change in one of the SOVO layers may initiate changes in the other layers. Figure 2 illustrates the levels of change, and their impact on the SOVO environment.

In general, our proposed framework classifies changes in the SOVO into:

- High level changes: represent changes in the business layer of the SOVO. This includes changes in the value network layer and the collaborative process layer.
- Low level changes: represent changes in the service layer of the SOVO. This includes changes in the SOVO service providers' layer.

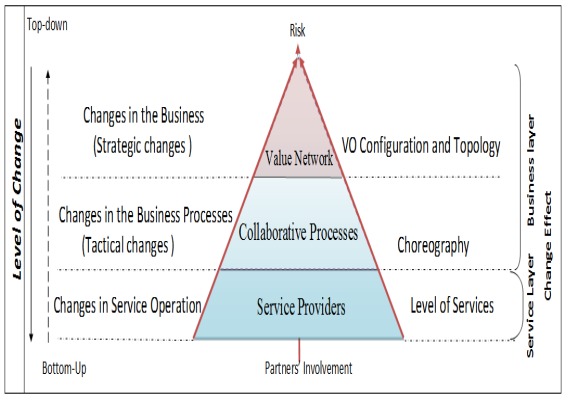


Figure 3: The levels of changes in the SOVO.

4.1.1 Value Network Layer Changes

The value network changes are high-level changes in the business processes, which affect the configuration and topology of the VO, and at some stages questions the feasibility of the VO itself. In the SOVO, the value network specifies the roles and interactions between different partners developed during the creation phase of the VO. Any changes in the value network are considered critical and require business decision-making and negotiation between the different VO partners involved.

Changes in the value network layer will affect

the VO's operational configuration, and topology. Changes in the VO's "operational configurations" are referring to the changes in the consortium of the legally independent organizations, their service contributions, value exchange and Service Level Agreement (SLA), which will impact the value added to the VO. In this layer, the value exchange among the SOVO partners can be any kind of service, money, product and information, which partners have agreed on during the creation and business opportunity discovery phase of the VO Lifecycle.

Changes in the VO "topology" are referring to the modification needed in the structure through which the VO partners interact and communicate with each other. The VO topology could be a supply chain topology where each partner communicates with its upper and lower neighbours; a star topology with a central authority where the collaboration between different partners is organized in star-like pattern between the central partner and other organizations; and a peer-to-peer topology where each partner is able to communicate and interact with the other partners without hierarchy or central authority (Danesh et al., 2011). Changes in the SOVO value network layer will cause changes in the collaborative process layer and service providers' layer.

4.1.2 Collaborative Process Layer Changes

Collaborative process changes are high-level changes in the business processes, which affect the choreography of the VO. Collaborative processes are high-level business processes which are defined to model the collaboration and integration between various organizations. The collaboration between different organizations in the SOVO is modelled using the service choreography, which focuses on collaboration and service interaction. It specifies each partner's role, activities and the sequences of service invocation.

Changes in the collaborative process layer means changes in the VO choreography. Changes in the choreography include changing in partners' roles, their assigned activities, and the service invocations. Changes in the collaborative process layer may or may not propagate to the changes in the value network layer (although, it might in the long-term) and require partners' negotiation and decision making. Furthermore, changes in the collaborative process layer will propagate to the lower level changes in the VO partners shared services.

4.1.3 Service Providers' Layer Changes

Service providers' layer changes are low-level changes in the VO service layer, which affect the level of the services provided by the VO partners. The VO partners can request a change in their VO shared services which affect the VO service providers' layer. The VO service providers can initiate *functional* and *non-functional* changes to their shared services. Functional changes refer to the changes that affect the functionalities of the service shared in the VO. It includes changes in the shared Web Services provided by the VO partners. These changes are initiated by the service providers, and affect the VO in the service providers' layer. Functional changes can be classified into structural and behavioural changes. Structural changes refer to the changes in the operational aspect of the shared services, whereas, the behavioural changes refer to the service providers' changes in its interactions with the VO in the collaborative process layer, or with the other service providers. Functional changes in service providers' layer occur when the web service description (WSDL) is modified (Akram et al., 2010).

Non-functional changes refer to the changes that are triggered from the VO partners' performance prospective. The privacy, security, reputation and availability are some of the performance measurements that service providers can consider in initiating changes to their communication and interactions with the VO or other service providers.

The involvement of the VO partners in the process of change increase as we move from the low-level changes to the high-level changes. This means, the number of partners involved in the change decision making and problem solving process will increase going from low level change to more complicated high level changes. Furthermore, the VO itself will face the risk of dissolution and the question of its feasibility as we move from the low-level changes to the high-level ones. If the VO participants were not able to collaborate with each other to solve the problem effectively, the VO will eventually dissolve.

4.2 The Triggers of Changes

The fact that the VO has distributed and independent partners causes changes in the SOVO occur more frequent. Changes can be initiated in different levels of change during the operation and evolution phase of the VO. These changes can be triggered for different reasons by the VO coordinator, change

management team or even by the VO partners themselves. Figure 3 illustrates different reasons which could lead to the change in the VO at the different levels.

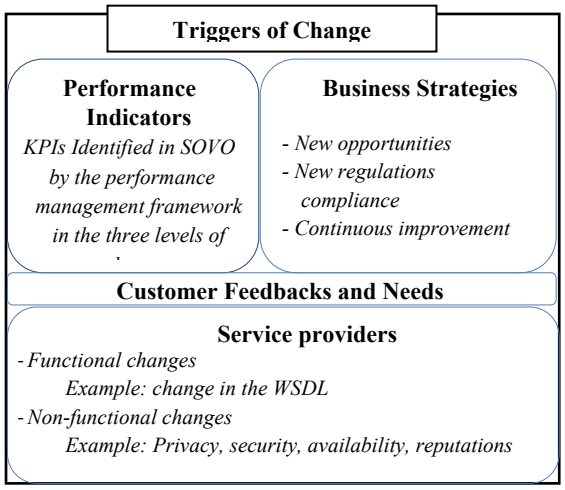


Figure 4: Triggers of change in the SOVO.

The performance indicators in the SOVO are defined in the performance management framework in the three levels of changes. Each level has its own key performance indicators (KPIs) to measure the performance of operation at that level. The VO performance management framework identifies the performance measurement criteria and their threshold which automatically trigger a change in one of the three layers (Kamali et al., 2012). This trigger of change is an internal trigger to the SOVO, based on its performance.

Customer feedbacks and needs can automatically trigger a change. The customers' feedbacks measure the external performance of the VO, and call for the future improvements in the VO services. By categorizing customers' feedbacks, and applying a predefined threshold, an event will be triggered if the customers' negative feedback passes the threshold. This external trigger will alert the SOVO change management team to investigate the problem, which may lead to the changes in the VO

Business strategies can trigger changes in the VO to meet new goals, objectives, or to add more values to the consortium of the independent organizations. The rapidly changing environment that the VO operates in forces it to adopt new strategies to achieve sustainable competitive advantage. The VO coordinator or the VO change management team can manually trigger changes in the VO based on the business strategies agreed on. Changes in the environment, markets, new business goals and shift in the VO operation force the business to trigger changes to meet their new strategies and objectives. Adopting a new VO business strategy can be triggered due to internal or external forces.

Service Providers can manually trigger changes to their shared services in the VO for functional or non-functional changes. A service provider can trigger a change in the VO by triggering a request for change (RFC) on his/her shared service. A service provider may trigger an RFC to update his service, which is being shared in the VO by adding new functionality or modifying the properties of the service.

5 EXAMPLE SCENARIOS OF CHANGE

The following scenarios demonstrate various triggers of changes at different levels of change. The VOWireless identified earlier is considered to demonstrate the scenarios.

Starting with the performance indicator as a trigger of change, assume that the "availability" of the VOWireless cell phone service provider does not pass the SLA threshold. This triggers an event and alerts the change management team. The problem then needs to be investigated by identifying its root cause. A possible solution could be substituting the cell phone service provider with a new one who is able to meet the agreed level of "availability". This is a change in the value network layer which will affect the VO topology and configuration, and will cause changes in the collaborative process layer, affecting the VO choreography. Furthermore, it will cause low level changes, affecting the service provider level. This change requires negotiation and business decision making between the VO participating organizations. This change needs a new VOWireless value network, new service choreography and new cell phone service provider.

Next, considering the business strategies as a trigger of change in the SOVO, VOWireless may initiate changes to add more values to their customers by adding a shipping service provider to the VO. The shipping service provider handles shipping the final product to the customers. This is a high level change in the value network which involves adding a new service provider to the VO. It will affect the VOWireless value network configurations and topology, and will cause changes in the collaborative process and service providers' layer. The change requires a new value network as

illustrated in Figure 5, new service choreography derived from the new value network, and a shipping service provider.

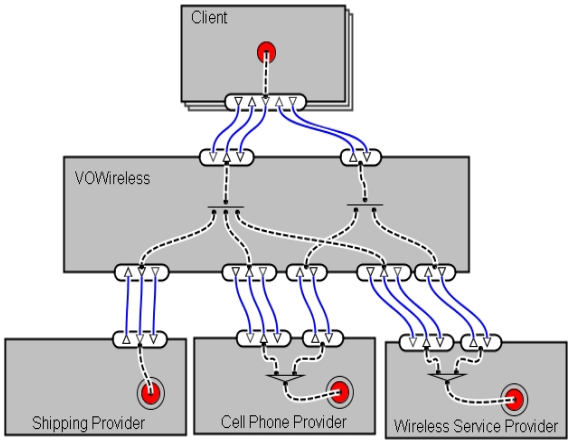


Figure 5: The modified VOWireless new value network.

Finally, considering the wireless service provider can trigger a change in VOWireless to change the shared service by updating the WSDL file that describes the service, and the way it is called. This is a low level change in the service providers' layer. Since it will not change the functionality (behavioural) of the service, it will not propagate to the higher levels of change in the VO. The wireless service provider can initiate the change by updating its shared service, and publish it again to be shared in the VOWireless.

6 CONCLUSIONS

The rapidly increasing number of web services, combined with the idea of virtual organizations (VO) to offer new services in a dynamic collaborative environment, introduces new challenges in their change managements. The VO members must constantly monitor the changes in their environment, and react to them in order to meet the market and customers' needs through improved services. With prior knowledge of the impacts, consequences and categories of the changes, combined with a predefined change processes, the VO participating organizations can respond quickly and flexibly to market demands to remain profitable and competitive.

We proposed a framework that guild the VO participants in identifying the impact, categories, and triggers of change in the SOVO environment. We specified a structural framework which include the level and triggers of change in the SOVO. The levels of change identified the SOVO three layers of change which include the value network layer, the collaborative process layer and the service providers' layer. The trigger of change identified different internal and external reasons which lead to change in the VO. This include the VO performance indicators, the customer feedbacks and needs, The VO business strategies and the service provider changes We presented different change scenarios based on the identified triggers and levels of change. The presented scenarios demonstrate how the VO and its partners can use the change management structural framework to understand the impacts, categories, triggers and severity of changes on the VO operations and collaborations.

Our next research plan is to identify and validate a change management *procedural framework* which provides a process of change based on the Information Technology Infrastructure Library (ITILV3), Engineering Change Management (ECM) and European Collaborative networked Organisations LEADership (ECOLEAD) best practices, recommendations and reference models, customized for the SOVO environment's change requirements.

REFERENCES

Abuelmaatti, A. & Rezgui, Y., 2008. Virtual Organizations in Practice: A European Perspective. In *Proceedings of the 14th Americas Conference on Information Systems (AMCIS)*. p. 142.

Akram, M. S., 2005. *Managing Changes to Service Oriented Enterprises*. State University.

Akram, S. et al., 2010. A change management framework for service oriented enterprises. *the international journal of next-generation computing*, 1(1).

Bianco, P., Kotermanski, R. & Merson, P., 2007. *Evaluating a Service-Oriented Architecture*,

Camarinha-Matos, L. M., Afsarmanesh, H. & Ollus, M., 2008. ECOLEAD and CNO base concepts. In *Methods and Tools for collaborative networked organizations*. Springer, pp. 3–32.

Danesh, M. H., Raahemi, B. & Kamali, M. A., 2011. A framework for process management in service oriented virtual organizations. In *Proceedings of the 7th International Conference on Next Generation Web Services Practices*. IEEE, pp. 12–17.

Drissen-silva, M. V & Rabelo, R.J., 2008. A model for dynamic generation of collaborative decision protocols for managing the evolution of virtual enterprises. In *Innovation in Manufacturing Networks IFIP – The International Federation for Information Processing*. Springer, pp. 105–114.

Drissen-Silva, M. V. & Rabelo, R. J., 2009. A collaborative decision support framework for managing the evolution of virtual enterprises. *The International Journal of Production Research*, 47(17), pp.4833–4854.

Dumitraş, T. et al., 2007. Ecotopia: An ecological framework for change management in distributed systems. *Architecting Dependable Systems IV Lecture Notes in Computer Science*, 4615, pp.262–286.

Erl, T., 2005. *Service-oriented architecture: concepts, technology, and design*, Prentice Hall.

Holley, K. & Arsanjani, A., 2010. *100 SOA Questions: Asked and Answered*, Prentice Hall.

Jansson, K. et al., 2008. Governance and management of virtual organizations. In *Methods and Tools for collaborative networked organizations*. pp. 221–238.

Kamali, S. M. A. et al., 2012. A framework for performance measurement in service oriented virtual organizations a value network approach to collaborative performance measurement. In *Proceedings of the 2010 International Conference on e-Business*. pp. 263 – 271.

Karvonen, I. et al., 2004. Challenges in the management of virtual organizations. *Virtual Enterprises and Collaborative Networks- IFIP International Federation for Information Processing*, 149, pp.255–264.

Liu, X. & Bouguettaya, A., 2007. Reacting to functional changes in service-oriented enterprises. In *Proceedings of of the International Conference on Collaborative Computing: Networking, Applications and Worksharing*. IEEE, pp. 264–270.

Negretto, U. et al., 2008. Vo Management Solutions VO Management e-Services. In *Methods and Tools for collaborative networked organizations*. Springer, pp. 257–274.

Obidallah, W. et al., 2013. Service oriented virtual organizations: A service change management perspective. In *Proceedings of the 26th IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*. IEEE, pp. 0–5.

OGC, 2007. *Service transition*, United Kingdom:: TSO (The Stationery Office).

Ul Haq, I. & Schikuta, E., 2010. Aggregation patterns of service level agreements. In *Proceedings of the 8th International Conference on Frontiers of Information Technology - FIT '10*. New York, New York, USA: ACM Press, pp. 1–6.

Wang, Y., Yang, J. & Zhao, W., 2010. Managing Changes for Service Based Business Processes. In *Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference*. Ieee, pp. 75–82.

An Investigation of Factors and Measurements for Successful e-Commerce Websites

Xiaosong Li, Wei Lian and Abdolhossein Sarrafzadeh
Department of Computing & Information Technology, Unitec Institute of Technology, Auckland, New Zealand
xli@gw.unitec.ac.nz, lianwei2006006@yahoo.com, hsarrafzadeh@gw.unitec.ac.nz

Keywords: e-Commerce, Website, Factor, Dimension, Measurement, Success.

Abstract: Business-to-Consumer (B2C) e-commerce is popular because of its convenience, speed and price. Although there has been intense debate about quality dimensions of e-commerce websites, more research is needed to find a well-established measurement. This empirical study identifies a set of measurements with 10 factors and their corresponding dimensions, including software development attributes based on the literature and the qualitative and quantitative data gathered from four different stakeholders. The survey results suggest that security, smooth transaction processes and smooth shopping processes are the most important concerns for online shoppers. The IS success model checking suggests that the proposed measurements are comprehensive. This work is compared with the customized ISO 9126 quality model.

1 INTRODUCTION

Business-to-Consumer (B2C) e-commerce is popular because of its convenience, speed and price (Luo et al., 2008). Although there has been debate about quality dimensions of e-commerce websites, however, they are not well-established, and more research is needed to find a measurement of e-commerce website success (Balfagih et al., 2010).

The existing relevant works contain either a broad framework which does not include specific and detailed information for more accurate, effective and practical evaluation of e-commerce websites, or the measurements focus only on certain aspects of the e-commerce websites, or the measurements are applicable to a general information system not specific for e-commerce websites. The existing DeLone & McLean IS Success Model was updated and extended by introducing new metrics for e-commerce system success measurement (DeLone & McLean, 2003). This framework was modified and extended further to be more specific on measuring e-commerce website success (DeLone & McLean, 2004). When these frameworks are applied to a specific website, further interpretations or refinements are required. An empirical study was conducted to develop a set of instruments of B2C e-commerce website quality (Cao et al., 2005). The instruments covered some aspects of service quality, such as responsiveness, information quality, and information accuracy. However, they omitted other aspects, such as page design and shopping process. The 7C framework was developed (Rayport & Jaworski, 2001) to evaluate e-commerce websites; however, it emphasizes the specific role of interface elements as a communication channel between retailers and their customers. The 7Cs evaluates an e-commerce website more from a user interface design perspective than information quality and usability perspectives. The 8C framework was developed based on the 7Cs to cover the Web 2.0 features (Yang et al., 2008). However, the scope of 7Cs was extended for generic web application, and the dimensions were interpreted more from technique implementation perspectives. A B2B e-commerce web application evaluation model was proposed in 2009 (Behkamal et al., 2009), in which a traditional software quality model, ISO 9126, was customised by identifying the B2B quality characteristics and adding the necessary attributes. The customisation was carried out in five steps. A numeric value was assigned to each of the quality characteristics from both the B2B application users' and developers' viewpoints. So the result emphasizes the user and development issues.

The long-term success is determined by their ability to generate positive net revenues (DeLone & McLean, 2004). Attracting Internet users, encouraging customer loyalty and improving service are aiming to increase profits ultimately (Bezes,

2009). These suggest that the ultimate measurement for e-commerce website success is profit. To help the business owners improve the quality of their e-commerce websites, there is a need to identify the factors that critically influence the quality of the websites and the measurements which can reflect the profit closely.

This empirical study focuses on B2C e-commerce websites that support shopping functionalities. It aims to find a set of comprehensive, specific and objective measurements based on the information possibly obtained from different sources for successful e-commerce websites. It uses the existing research findings as a starting point and captures the critical factors that contribute to the success of an e-commerce website from design, content, shopping process, security, management and technique perspectives. Ten important factors and their dimensions were identified initially (Lian, 2012). To verify and refine these factors, data were collected from an online survey of e-commerce website users, interviews of relevant professionals and online tool testing. The data were interpreted and analyzed with the help of two simple quantitative instruments: mean and standard deviation; and the factors as well as their dimensions were refined. The refined factors were examined against the six dimensions of the updated DeLone and McLean (2004) information systems success model. The refined factors were also examined against the conceptual model proposed by Ghandour group (Ghandour et al., 2008).

In the following sections, the initial measurements are described first, that is, followed by the data analysis and the measurement refinement, the proposed measurements are validated, and finally the result is discussed and the paper is summarized.

2 INITIAL FACTORS

On the basis of a study of the existing publications, this research identifies 10 factors that may contribute to the success of e-commerce websites. The detailed dimensions for each factor were also identified at the same time. Following are the 10 factors, their dimensions, their literature sources and their codes which will be referred to in the rest of this paper:

- ***Factor 1. Site Popularity (F1)*** *(Ghandour et al., 2008)*: Keeping high daily reach (F1D1), A large number of unique users (F1D2), A larger number of visitors (F1D3);
- ***Factor 2. Website Profit (F2)*** *(Carney, 2005; Barnes, 2002)*;
- ***Factor 3. Usability and Performance (F3)*** *(Haenlein, 2004; Kuballa, 2007; May, 2000; Chen, 2005)*: Speed of site load (F3D1), Website consistence for the leading browsers (F3D2), The home page is well-organized and easy to read and understand (F3D3); Logical and clear navigation choices with consistence and availability (F3D4); Easy to use navigation (F3D5); Easy to use search boxes (F3D6); Functionality (F3D7); Error message resolution (F3D8);
- ***Factor 4. Security and Privacy (F4)*** *(Gehling & Stankard, 2005)*: Escrow service (F4D1); Personal information privacy (F4D2); Secure server (F4D3); Secure payment mechanism (F4D4); Antivirus software (F4D5); Firewall (F4D6); Authentication (F4D7);
- ***Factor 5. Website Design (F5)*** *(Lau, 2007; Blanchard & Fabrycky, 1990)*: Use of animated gifts and ticker tapes (F5D1); A balance between design and functionality (F5D2); Page fitness for scrolling entire page (F5D3); Simple and logical web layout (F5D4); Use of colors (F5D5); Expression of company image (F5D6); Easy to understand product categories (F5D7); Appropriate typefaces and font sizes (F5D8); Ease in finding checkout button (F5D9); Shopping cart location (F5D10);
- ***Factor 6. Website Content (F6)*** *(Leea & Kozar, 2006)*: Quality information (F6D1); Good service (F6D2); Easy to understand (F6D3); Catalog information organized logically (F6D4);
- ***Factor 7. Transaction Process and Volume (F7)*** *(John, 2004)*: The site supports secure credit card transactions (F7D1); Different payment methods available (F7D2); User-friendly rules and policy for sellers and buyers (F7D3); Good trade volumes (F7D4);
- ***Factor 8. Shopping Process (F8)*** *(Spiekermann et al., 2001)*: Reliable delivery process (F8D1); Easy to obtain refunds (F8D2); Easy to use the shopping cart (F8D3); Easy to add or delete items from the shopping cart (F8D4); Easy to view the contents of the shopping cart (F8D5); An obvious click path to the cash register (F8D6); Accessible product inventory and availability of information (F8D7); Customer rating/recording system (F8D8);

- ***Factor 9. Technical Support (F9)*** *(Citera & Beauregard, 2005)*: Availability of site map (F9D1); Customer online help (F9D2); Discussion board (F9D3); User guide (F9D4); FAQs (F9D5); Contact details (F9D6);
- ***Factor 10. System Development (F10)*** *(Anil et al., 2003)*.

3 THE DATA AND ANALYSIS

Customer experience could influence customers' perceptions of value and service quality, therefore, affecting customer loyalty. This is critical for the success of an e-commerce website. An online survey (Lian, 2012) was used to solicit customers' opinions on most of the proposed factors, except F2 and F10, which are more appropriate for professionals' opinions. 52 e-commerce website users from a local technology institute participated in the survey, of which the majority are aged between 18 and 30 (Lian, 2012). These young tertiary students are active on the internet and have good computer skills. The results of the survey may reflect this situation. Eight selected professionals participated in the interviews. The chosen participants needed to understand and have an interest in the research area, as well as have certain knowledge of e-commerce websites. In addition, the participants needed to have the ability and insight to perceive critical success factors for e-commerce websites with a technical and business focus. Two academic researchers, three business managers and three e-commerce website developers were interviewed. The authors believe that they possessed the required characteristics and experiences.

3.1 Online Survey Data

The participants were asked to choose which particular factors are critical for e-commerce website success and how they weight each of the dimensions of a particular factor (Lian, 2012). The questions were presented using a five-point Likert-type scale, 5 being the most desirable, and 1 being the least. A dimension score is the average of all the scores for that dimension from all the participants. A factor score is the average of all the scores for that factor from all the participants. Table 1 is a summary of the results.

The standard deviation of the dimension scores for a factor was calculated to check the diversity of the dimensions in the participants' opinions. A small standard deviation may indicate that the dimensions proposed for that particular factor are highly correlated; a large one may indicate that certain dimensions are inappropriate.

Table 1: Online survey results.

Factors	Dimensions	Dimension Score	STDEV	Factor Score
F1	F1D1	3.79	0.093	4.21
	F1D2	3.84		
	F1D3	3.97		
F3	F3D1	4.03	0.139	4.22
	F3D2	3.94		
	F3D3	4.14		
	F3D4	4.09		
	F3D5	4.14		
	F3D6	4.39		
	F3D7	4.14		
	F3D8	3.97		
F4	F4D1	3.69	0.359	4.34
	F4D2	4.29		
	F4D3	4.26		
	F4D4	4.54		
	F4D5	4.23		
	F4D6	4.09		
	F4D7	4.37		
F5	F5D1	3.00	0.359	4.22
	F5D2	4.00		
	F5D3	3.57		
	F5D4	3.91		
	F5D5	3.56		
	F5D6	3.43		
	F5D7	4.17		
	F5D8	3.71		
	F5D9	4.00		
	F5D10	3.78		
F6	F6D1	3.97	0.075	4.29
	F6D2	4.11		
	F6D3	3.97		
	F6D4	4.09		
F7	F7D1	4.20	0.109	4.32
	F7D2	4.20		
	F7D3	4.14		
	F7D4	3.97		
F8	F8D1	4.09	0.135	4.29
	F8D2	4.40		
	F8D3	4.00		
	F8D4	4.23		
	F8D5	4.14		
	F8D6	4.11		
	F8D7	4.31		
	F8D8	4.06		
F9	F9D1	3.44	0.278	4.17
	F9D2	3.85		
	F9D3	3.56		
	F9D4	3.56		
	F9D5	3.71		
	F9D6	4.21		

The survey results show that F4 and F7 got the highest factor scores of 4.34 and 4.32, respectively. This suggests that security and smooth transaction processes are the two most important concerns for online shoppers.

System security becomes a more significant system-quality issue as e-commerce is typically conducted over the Internet (DeLone & McLean. 2004). Good protection on confidential information should help to gain the consumer's trust (Yazdanifard et al., 2011). An e-commerce website must provide a secure user interface system and an easy-to-use transaction process. F9 got the lowest score of 4.17, suggesting that customers prefer a convenient and easy-to-use shopping system than using a help system. This may also reflect that the survey samples are highly computing skilled. F8 scored 4.29 which is higher than F5 at 4.22. This suggests that the shopping process and functionality is more important to the customers than the site's visual appearance.

Table 2: The difference between the dimension average of a factor and its factor score.

Factors	Dimension Average	Factor Score	Difference
F1	3.87	4.21	0.34
F3	4.11	4.22	0.11
F4	4.21	4.34	0.13
F5	3.71	4.22	0.51
F6	4.04	4.29	0.25
F7	4.13	4.32	0.19
F8	4.17	4.29	0.12
F9	3.72	4.17	0.45

The scores of all the dimensions of a particular factor were averaged to compare with that factor's score to gauge the validity of the dimensions for that particular factor. Table 2 shows the dimension averages and the differences between the dimension average of a factor and its factor score. A large difference may indicate the omission of certain dimensions for that factor. For example, the largest difference between the two scores is 0.51 for F5, which suggests that the identified dimensions for this factor may not be suitable or complete.

Customization defined in the 7Cs (Rayport & Jaworski, 2001) was missed out in the dimensions of F5. With customization, a user is provided with functions that enable him/her to enter personal details or customize the look of a banner. So, the dimension "Be able to customize the look of the user interface" with code F5D11 is added to F5. F3 got a small average score difference of 0.11 and a relatively small standard deviation of 0.139, which suggests that the dimensions for this factor are well defined. For example, "Search facility" was highlighted in (Cao et al., 2005) and it was included in F3D6. The difference for F9 is 0.45, which is the second largest. There may be serious omissions for this factor. Service quality can be measured by the effectiveness of on-line support capabilities, such as answers to frequently asked questions or responsiveness in the e-commerce environment (DeLone & McLean, 2004). This is confirmed by (Cao et al., 2005), where responsiveness is a construct with 5 items. So "Responsive to customers' inquiries" with code F9D7 is added to F9. The difference for F6, 0.251, is large as well. This suggests omissions. The small standard deviation 0.075 suggests that the proposed dimensions may only focus on certain aspects of this factor. The dimensions for F6 show that some popular dimensions like "use video demonstration" (Cao et al., 2005) are missing, thus, it is added to code F6D5, and, similarly, "use customer testimonial" with F6D6 is added. For F7, the average score of F7D4 is much lower than the others, suggesting that F7D4 may not belong to this factor. Removing F7D4 from the factor reduces the difference to 0.14, which is more appropriate. The name of F7 has been changed, therefore, to ***Transaction Process.***

3.2 Interview Data

Each group of professionals was given a slightly different set of questions to reflect their expertise (Lian, 2012). While they shared some opinions with the online survey participants, there were a few differences that need to be highlighted.

Only the professionals' opinions were solicited for F2 and F10. For F2, six out of eight professionals thought that website profit is a success factor and no meaningful comments were given. On the other hand, Transaction Volume was fully supported with strongly supportive comments, such as, "there must be a problem if the trading volume is too small", "transaction volume and website success are positively correlated", and "transaction volume reflects the commercial value of a website". In many cases, commercial transaction volume is positively related to commercial profit. It is reasonable to add "Good transaction volume" with code F2D1 as a dimension of F2.

F10 received full support as well, with many meaningful comments. The comments added desirable attributes for e-commerce websites from a software development perspective. Here is a

summary of the comments, which define the dimensions for this factor: System Requirements-Reliability (F10D1); System Requirements-Usability (F10D2); System Requirements-Data Integrity (F10D3); Resources-Technologies and Tools (F10D4); Resources-Network & Internet Access (F10D5); Resources-Relevant Knowledge (F10D6); Process-Back up (F10D7); Process-Documentation (F10D8); Process-Methodology (F10D9).

3.3 Online Testing Data

The online testing tool Alexa was used to test 30 e-commerce websites, including three very successful sites, TaoBao, eBay and Amazon (Lian, 2012). The results showed that TaoBao, eBay and Amazon got a much higher Alexa Traffic Rank than the others. They scored 15, 24 and 17, respectively, while the highest rank for the others was 387. It is reasonable to use Alexa Traffic Rank (Global) as a quantitative measurement for e-commerce website success. The average score difference for F1 is quite large, 0.34, with a small standard deviation, 0.093. This suggests omissions again. So "Traffic rank" with code F1D4 is added to F1. Table 3 is a summary of the revised factors and their dimensions.

Table 3: Revised factors and their dimensions.

Factors	Dimensions
F1	F1D1, F1D2, F1D3, F1D4
F2	F2D1
F3	F3D1, F3D2, F3D3, F3D4, F3D5, F3D6, F3D7, F3D8
F4	F4D1, F4D2, F4D3, F4D4, F4D5, F4D6, F4D7
F5	F5D1, F5D2, F5D3, F5D4, F5D5, F5D6, F5D7, F5D8, F5D9, F5D10, F5D11
F6	F6D1, F6D2, F6D3, F6D4, F6D5, F6D6
F7	F7D1, F7D2, F7D3
F8	F8D1, F8D2, F8D3, F8D4, F8D5, F8D6, F8D7, F8D8
F9	F9D1, F9D2, F9D3, F9D4, F9D5, F9D6, F9D7
F10	F10D1, F10D2, F10D3, F10D4, F10D5, F10D6, F10D7, F10D8, F10D9

4 INFORMATION SYSTEM SUCCESS MODEL

According to DeLone and McLean (2004), all of the proposed measures for e-commerce website could be classified under six dimensions. To validate the proposed measurements in Table 3, it is helpful to check them against the six dimensions of the DeLone & McLean Model. Table 4 shows detailed information for each dimension.

For B2C e-commerce websites that support shopping functions, the main stakeholders are the customers and the business owners. For the net benefits, we should consider the benefits for both of them. When the customers find the site convenient or are happy with the quality/price of the products they purchase, so the website can make a profit. Website Profit (F2) can be used to measure this. Site Popularity (F1) can be used to measure customer satisfaction.

Table 4 shows the proposed factors' classification in the DeLone & McLean Model. The factors are distributed into all six dimensions, so they should be comprehensive; however, they are not necessarily accurate. For example, F1 can reflect customers' satisfaction to a certain extent, however, possibly not completely.

Table 4: Checking with DeLone & McLean Model.

Dimensions	Measures	Factors
System quality	Measures the desired characteristics such as usability and response time	F3, F10
Information quality	Captures the content issue such as personalization, complete, relevant, easy to understand and secure	F4, F5, F6
Service quality	The overall support delivered	F9
Usage	Measures everything from a visit to a website and navigation to information retrieval	F7, F8
User satisfaction	Measures customers' opinions	F1
Net benefits	Captures the balance of the positive and negative impacts	F2

In the DeLone & McLean Model, the role of computing was not strongly suggested nor recommended. This was considered in an extended model proposed by Ghandour et al. (2008), where there were three main phases of e-commerce website success: development, use and consequences of the website. This study categorized the factors proposed in Table 3 by collaborating with the interviewed professionals. They were asked to classify the ten factors in the three given categories (Lian, 2012). The results showed that 6 factors (F3, F4, F5, F6, F9 and F10) belong to the development; 2 factors (F7 and F8) belong to the use; and 2 factors (F1 and F2) belong to the impact. Each category got more than half of the experts' support. So the proposed

measurements fit in the extended model. It also provided empirical evidence support for the extended model.

5 DISCUSSION AND SUMMARY

A set of measurements with 10 factors and their corresponding dimensions was identified in Table 3 based on the literature and the qualitative and quantitative data gathered from four different e-commerce website stakeholders. It is the combination of these factors that measures the success of an e-commerce website rather than an individual factor. The survey results suggested that security, smooth transaction processes and smooth shopping processes are the most important concerns for online shoppers. Good protection on confidential information should help to gain the consumer's trust (Yazdanifard et al., 2011) and maintain customer loyalty. The IS success model (DeLone & McLean, 2004) checking suggested that the proposed measurements are comprehensive; however, they are not necessarily accurate. The proposed measurements fit in the extended model (Ghandour et al., 2008), which also provides empirical evidence support for this model. To refine the measurements further, empirical tests involving a set of e-commerce websites and their customers/owners are required.

Comparing the proposed measurements with the existing work, the customized ISO 9126 quality model (Behkamal et al., 2009) seems the closest. They both define the quality factors in a hierarchical structure and assign numerical values to the quality factors. However, one is for B2B e-commerce websites and the other is for B2C. In terms of quality attributes, the customized ISO 9126 quality model emphasizes the user and development factors, such as “Functionality”, “Reliability” and “Usability”; the proposed measurements emphasize the customer and business-related factors more, such as “Website Design”, “Website Content”, “Shopping Process” and “Transaction Process and Volume”. In terms of process, five steps were defined in the customized ISO 9126 quality model, particularly in step 4, a widely-used method, analytical hierarchy process (AHP) is carefully followed. This systematic approach makes it easy to generalise the five steps for other types of e-commerce websites such as B2C. However, the factors should be revised to include more customer-related features. On the other hand, this work used the scores given by the users and adjusts the quality factors by means of simpe statistics instruments. These were further improved by considering the opinions of the other stakeholders. In terms of data, only users’ viewpoints and developers’ viewpoints were considered in the customized ISO 9126 quality model; in the proposed measurements, users’, developers’, business owners’ and researchers’ viewpoints were considered, therefore, the resulting quality factors should be more complete.

The small sample size of the online survey limited the use of rigorous statistics in the data analysis, instead, a mixed approach using combined qualitative and quantitative analysis was used on these data. A large sample with a more general population should improve the accuracy of the results.

REFERENCES

Anil, S., Ting, L. T., Moe, L. H. & Jonathan, G. P. G., 2003. Overcoming barriers to the successful adoption of electronic commerce in Singapore. *International Journal of Mobile Communications*, 1(1-2), pp. 194-231.

Balfagih, Z., Mohamed, N. & Mahmud, M., 2010. A Framework for Quality Assurance of Electronic Commerce Websites. *Chapter 9, E-Commerce*, Edited by Kyeong Kang, ISBN 978-953-7619-98-5, Published by In-tech.

Barnes, S. J., 2002. The electronic commerce value chain: Analysis and future developments. *International Journal of Information Management*, 22(2), pp. 91-108.

Behkamal, B., Kahani, M., & Akbari, M. K., 2009. Customizing ISO 9126 quality model for evaluation of B2B applications. *Information and Software Technology*, 51(3), pp. 599-609.

Bezes, C., 2009. E-Commerce Website Evaluation: A Critical Review. Working Paper.

Blanchard, B. & Fabrycky, W., 1990. *Systems engineering and analysis*. Cambridge, UK: Cambridge University Press.

Cao, M., Zhang, Q. & Seydel, J., 2005. B2C e-commerce web site quality: an empirical examination. *Industrial Management & Data Systems*, 105(5), pp. 645-661.

Carney, M., 2005. Trade Me Success Secrets: How to Buy Better and Sell More Profitably on New Zealand's Most Popular Auction Site. Buckinghamshire, UK: Activity Press.

Chen, R., 2005. Modeling of User Acceptance of Consumer E-Commerce Website. *Lecture Notes in Computer Science*, 38(6), pp. 452-462.

Citera, M. & Beauregard, R., 2005. An experimental study of credibility in e-Negotiations. *Psychology and Marketing*, 3(12), pp. 310-332.

DeLone, W. H. & McLean, E. R., 2003. The DeLone and McLean Model of Information Systems Success, A

Ten-Year Update. *Journal of Management Information Systems*, 19(4), pp. 9-30.

DeLone, W. H. & McLean, E. R. 2004. Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *International Journal of Electronic Commerce*, 9(1), pp. 31-47.

Gehling, B. & Stankard, D., 2005. eCommerce security. In *Proceedings of the 2nd Annual Conference on Information Security Curriculum Development*, pp. 331-340. NY, USA: ACM.

Ghandour, A., Deans, J., Benwell, G. & Pillai, P., 2008. Measuring eCommerce Website Success. In *Proceedings of the 19th Australasian Conference on Information Systems*, Christchurch, New Zealand: ACIS.

Haenlein, M., 2004. *An Exploratory Investigation of E-business Success Factors Using Partial Least Squares Analysis*. Perlin, Germany: Cuvillier Verlag.

John, M., 2004. Can Internet services facilitate commerce? Findings from the Greek telecommunications market. *International Journal of Mobile Communications*, 2(2), pp. 188-198.

Kuballa, J., 2007. *Key Factors of Successful E-commerce*. Paris, France: GRIN Verla.

Lau, R., 2007. Towards a web services and intelligent agents-based negotiation system for B2B eCommerce. *Electronic Commerce Research and Applications*, 6(3), pp. 260-273.

Leea, Y., & Kozar, K. A., 2006. Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach. *Decision Support Systems*, 42(3), pp. 1383-1401.

Lian, W., 2012. Success Factors for Transactional eCommerce Websites: an Investigation Focusing on Technical and Business contexts. Master's Thesis, Unitec Institute of Technology.

Luo, X., Tu, Y., Tang, J. & Kwong, C. K., 2008. Optimizing customer's selection for configurable product in B2C e-commerce application. *Computers in Industry*, 59(8), pp. 767–776, October.

May, P., 2000. *The business of ecommerce: from corporate strategy to technology*. Cambridge, UK: Cambridge University Press.

Rayport, J. E. & Jaworski, B. J. 2001. *E-Commerce*, McGraw-Hill Higher Education, New York.

Spiekermann, S., Grossklags, J., & Berendt, B., 2001. E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. In the *Proceedings of the 3rd ACM conference on Electronic Commerce*, New York: ACM.

Yang, T. A., Kim, D. J., Dhalwani, V. & Vu, T. K., 2008. The 8C Framework as a Reference Model for Collaborative Value Webs in the Context of Web 2.0. In *the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, p. 319.

Yazdanifard, R., AbuTabik, M. & Seyedi, A., 2011. Security and Trust in Electronic Commerce - Finding the Safe Side, In *the Proceedings of the International Conference on Information Communication and Management*, IPCSIT vol.16, IACSIT Press, Singapore, pp. 141-146.

Who Runs the Internet?
Classifying Autonomous Systems into Industries

Annika Baumann and Benjamin Fabian
Institute of Information Systems, Humboldt University Berlin, Spandauer Str. 1, 10178 Berlin, Germany
{annika.baumann, bfabian}@wiwi.hu-berlin.de

Keywords: Internet, Autonomous System, Industry, Classification.

Abstract: The Internet consists of a network of Autonomous Systems (ASs). To understand which kind of organizations control those ASs can help to better assess the Internet structure in terms of economic interests and reliability. The current paper proposes a novel classification approach by combining AS-specific data with business data from the United States Securities and Exchange Commission. Furthermore, more detailed industry classes than in previous works are considered, inspired by the North American Industry Classification System (NAICS). Using our methodology on a recent data set, we were able to classify 56.69 % of the considered ASs into industries. This lays a foundation for our future work on investigating the important players of the Internet backbone as well as their economic interests and risks.

1 INTRODUCTION

The Internet expanded rapidly during the last decade. From 2001 to 2013, the fraction of the world population using the Internet increased from 8.0 % (International Telecommunication Union, 2011) to an estimated 38.8 % (International Telecommunication Union, 2013), with a simultaneous population increase from 6.1 billion (United Nations Population Fund, 2001) to 7.2 billion people (U.S. Department of State, 2013), resulting in approximately 2.8 billion Internet users today versus 0.49 billion in 2001. This rapid growth in users resulted in a heterogeneous and complex system, making analysis and modelling of the Internet structure difficult.

Our paper is part of an on-going research project that is investigating how the Internet of today is structured in terms of economic interests, control and reliability. Who are the important players of the Internet backbone, what are their economic interests and risks with respect to their business models, and what are the implications for reliability, security and privacy as well as political control?
Our first step towards approaching these goals is to classify the important organisations that control Autonomous Systems (ASs) of the Internet according to business categories, which could support future analyses along all of those dimensions. For example, with respect to reliability and security, common methods assess the robustness of the Internet structure based on graphs and modelling the Internet as an abstract complex network consisting of nodes (each representing an AS) that are connected via edges. However, such approaches solely focus on topology-based robustness and so far ignore the highly economically driven character of the Internet, as well as corresponding heterogeneous risks of attack and control.

At an organizational and global routing level of abstraction, the Internet can be considered as composed of ASs. An AS can be defined as "a group of IP networks run by one or more network operators with a single clearly defined routing policy; when exchanging routing information to the outside, each AS is identified by a unique number (Réseaux IP Européens, 2011). The *Internet Corporation for Assigned Names and Numbers* (ICANN) and, via delegation, the *Regional Internet Registries* (RIR) are responsible for registration of these AS numbers (ASNs). The amount of registered ASs increased from roughly 10,000 in the year 2000 to more than 60,000 in 2013 (Potaroo, 2012), which is also another indicator for the substantial increase of Internet complexity.

Classifying the major players of the Internet backbone is an interesting challenge in itself because publicly available business data is sparse. Our approach presented in this article focuses on

analysing the public registration information for AS numbers. Moreover, we present an approach for the classification of ASs into detailed industry classes in order to better understand the organizational and economic patterns of the Internet.

The rest of the paper is structured as follows: Section 2 discusses related work. Section 3 presents the data sources, followed by Section 4 on our methodology. Section 5 presents our results, and Section 6 concludes the paper.

2 RELATED WORK

Some earlier research articles proposed approaches for classifying ASs into various categories. The classification approach used in our paper was initially inspired by the methods employed by Dimitropoulos et al. (2005). Based on an expert system that uses text classification techniques, the authors used organization names to categorize ASs. Each AS was assigned to one or more of the basic classes Internet service providers (ISP), Internet exchange points (IXP), network information centers (NIC), companies providing no Internet service as well as education- and research-, military-, government- and health-related networks. The authors were able to classify 20,598 out of 32,689 ASs in 2005, which corresponds to 63.01%.

Another work (Dimitropoulos et al., 2006) used even more coarse-grained classification categories, namely only large and small ISPs, customer ASs, universities, IXPs and NICs. The method applied was based on the AdaBoost algorithm (Freund and Schapire, 1997) using several attributes (e.g., organization description; number of inferred providers, customers and peers; number of advertised IP prefixes) to classify the relevant ASs into their respective classes. The authors were able to classify 95.3% of 19,537 ASs with an accuracy of 78.1%.

The main focus of the work by Chang et al. (2005) was to estimate traffic volume between individual ASs. For this, the authors classified ASs regarding their initial utility, which resulted into the three classes web hosting, residential access and business access. The methodology used by the authors is different from other work conducted in this area. Instead of investigating an individual AS and assigning it to a class, they created a class and tried to find relevant ASs on the Internet. The authors were able to identify 56% of all BGP-advertised ASs with their approach.

The primary focus of the paper by Dhamdhere and Dovrolis (2011) was to analyse the evolution of the AS ecosystem over the last 12 years. ASs were classified into the classes enterprise customers, small and large transit providers, access/hosting providers and content providers. A decision tree approach was applied for classification. In order to build the training set, for each class 50 ASs were classified manually. Afterwards, the classification was conducted for 42,000 ASs by using the number of customers and the number of peers as independent variables. Classification accuracy for the classes ranged between 76% and 82%.

All of those articles have in common that the proposed classes are not comprehensive and do not resemble real industries. Thus they contribute not much to a better understanding of the industry structure behind the ASs comprising the Internet. Our work addresses this research gap by proposing a classification approach that adopts fine-grained industry classes.

3 DATA SOURCES

3.1 CAIDA

The Cooperative Association for Internet Data Analysis (CAIDA) “is a collaborative undertaking among organizations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure.” (CAIDA, 2011). One project offered by CAIDA is the *AS Rank* project (CAIDA, 2012). It is based on Border Gateway Protocol (BGP) routing data collected by RouteViews (2013) and the RIPE NCC (2013). The list of ASs that is used in our paper contains the information of 59,576 ASs. An excerpt of the dataset can be seen in Figure 1. For the purpose of classifying ASs into industry classes mainly the *org name* attribute was considered as highly relevant.

```
# format: AS number|source|AS name|country|org name|org_id|date
1|ARIN|LVLT-1|US|Level 3 Communications|LVLT-ARIN|20120224
```

Figure 1: Excerpt of CAIDA *AS Rank*.

3.2 SEC

The U.S. Securities and Exchange Commission (SEC) is a government agency in the USA (United States Securities and Exchange Commission, 2013). Its primary purpose is to regulate securities and

enforce federal securities laws. Every company publicly traded in the United States has to file certain documents with the SEC. The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system makes those filings available to the public. This can be used to gather the Standard Industrial Classification (SIC) code for the company (Figure 2). An SIC code can be directly mapped to an NAICS code using a mapping table (CareerOneStop U.S., 2013). Thus it is possible to uniquely identify the industry of an AS's organization by use of the EDGAR system. A limitation is that only organizations that are listed on a stock exchange in the USA can be found in the system.



Figure 2: Excerpt of SEC EDGAR result.

3.3 RIR as Information

As an additional information source, data from the RIRs was retrieved. The website *cidr-report.org* contains AS information from all RIRs. It allows searching for individual ASs and returns the information that comes from the WHOIS services of the individual RIRs. In order to simplify the data retrieval process, this website was also used to retrieve AS-specific WHOIS information instead of using the WHOIS services of different RIRs. A sample of such information can be seen in Figure 3.

```
aut-num:        AS6619
as-name:        SAMSUNGSDS-AS-KR
descr:          SamsungSDS Inc.
descr:          Seoul Yeoksam-dong Gangnam-gu   707-19
descr:          135-080
country:        KR
```

Figure 3: Sample of RIR AS information.

4 METHODOLOGY

Figure 4 gives an overview of the process of classifying the ASs presented in this paper. As a first step, the relevant industry classes for the classification approach needed to be defined. Their definition draws from the North American Industry Classification System (NAICS, 2013). Due to the intrinsic online setting of our investigation, special adjustment was necessary, meaning that several of these classes were either merged, dropped or changed. In the case of ASs, some industries are missing at all while some of them are overrepresented. Therefore, the NAICS was only used as a basis for the classification approach in our particular setting. An overview of the classes can be found in the Appendix.

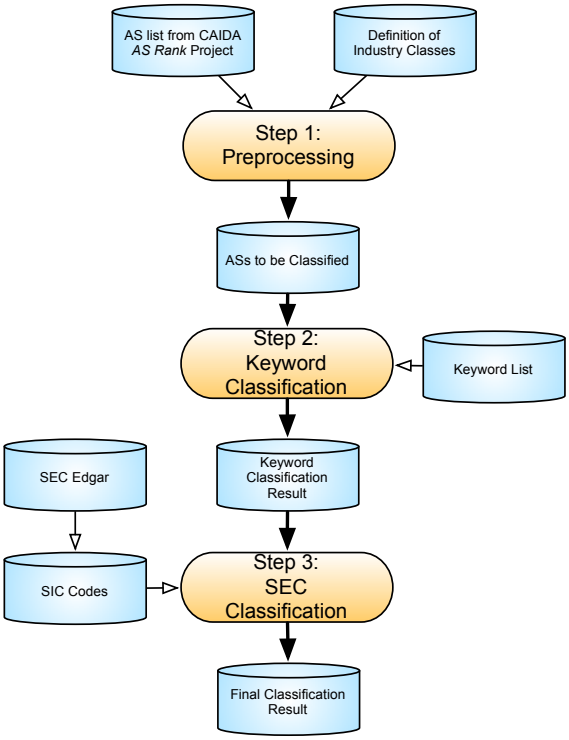


Figure 4: Process of AS industry classification.

Step 1: Preprocessing. The initial AS list included data from the year 2012 and was taken from the CAIDA AS Rank project (CAIDA, 2012); it contained 59,576 ASs. In order to only include reasonable and recent data, the list was preprocessed. At first, the information gathered from the RIRs was used to filter for inactive ASs. This reduced the list by 17,830 ASs, leaving 41,746 ASs to classify. Furthermore, all ASs that did not have an according organization name, i.e., all entries either containing no specification of the underlying organization or being a no registry entry, were removed from the list. Eliminating 1,362 ASs, this step left 40,384 ASs in the list.

Step 2: Keyword Classification. In the next step, a keyword list was created by analysing word and phrase frequencies with the help of an occurrence counting of words, bi-grams and tri-grams. All words and phrases that appeared quite frequently were analysed in more detail. It was assumed that tri-grams needed to occur at least five times, bi-grams ten times and simple words twenty times to be selected for deeper analysis. The rationale behind

this procedure was to include only those words and phrases that are most frequent and therefore important. This makes it possible to classify several ASs at the same time based on a single phrase or keyword. Keywords were mainly defined in such a way that the organization name or a part of it had to comply with the complete keyword. This means that for example in case of the keyword "ship" only the word itself would fit and not "membership" or "ownership". This was done to ensure the reliability of keywords by avoiding undesired mismatches. The selection of keywords itself was randomly cross-checked based on real data to further ensure their reliability and unambiguity. Only those words or phrases were chosen whose unambiguity in relation to industry classification was satisfactory. For example the keyword "Internet service provider" is highly reliable if it comes to sorting into the category ISPs & Networks, while "service provider" might lead to wrong results for the same category. Organizations having a (part of their) name such as "content service provider" would also fit into such a category.

In order to minimize wrong categorizations, an iterative learning process was applied. The procedure was as follows: based on the first selection of keywords, the AS numbers were categorized into the industry classes created so far. Each category was then checked for wrong categorizations. For this purpose, the list of categorized ASs and their underlying organization was reviewed manually. If the categorization of an ASN was wrong, the reason was identified and eliminated with the help of refined or discarded keywords. This procedure helped to ensure that only those keywords remained that are at the same time reliable and general. In order to check for further yet not identified keywords, a list was generated that contained all non-categorized ASNs. This list was then manually checked for further keywords at each iteration. This was particularly important in case of misspelling and language-specific variations. For example, the keyword "university" was represented by many language specific variations such as "universitas", "universidad" or "univ". An example for misspelling is "network *infomation* center" which occurred at least seven times in the list. Such variations were additionally included in the keyword list for each category.

Based on this extended and refined keyword list, the procedure started from the beginning and was repeated again. The complete list of the industry classes created and their respective definition are shown in Figure 7. The keywords used for each industry class are given in the Appendix.



Figure 5: Ambiguous EDGAR search result for "Sprint".

Step 3. SEC Classification. A Java program was written to download information from the SEC EDGAR system. The organization name was used to search for the company. For 40,384 search requests, 2,732 entries could be found in the EDGAR system. However, sometimes the same company has several names, which resulted in more than one outcome for the organization name. An example of such an ambiguity can be found in Figure 5. Because there was no reliable way to uniquely identify the correct entry in such a case automatically, all entries with multiple search results were eliminated which led to 1,706 remaining search results. Furthermore some companies had no SIC code and were eliminated as well. This resulted in 469 ASs that could additionally be classified into industry groups.

5 RESULTS

5.1 Keyword Classification

Applying the method described above and using the keywords shown in the Appendix to classify the 40,384 ASs, resulted in 22,786 or 56.42 % of classified ASs. The industry class distribution based on keyword classification only can be seen in Figure 6. According to this data, most frequently the organizations that own ASs belong to the industry classes *Education & Research*, *Finance & Insurance*, *ISPs & Networks*, and *Telephone & Communications*. This class distribution seems to be intuitive: ISPs, telephone and IT companies as well as universities have more incentives to register an AS than for example a travel agency because ASs classified into these categories are often related to communications, but often also represent major institutions that have a high tendency to own an AS simply because of their size.

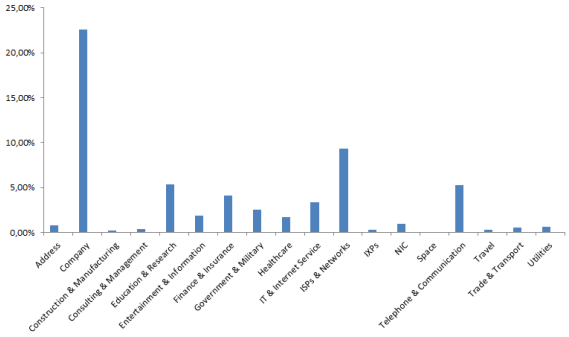


Figure 6: Industry class percentages based on keyword classification.

Of all clusters (apart from the generic *Company* cluster) *ISPs & Networks* is the category that is most frequent and accounts in case of both classification approaches for around ten percent of all the classified entities. This is an expected result since ASs pertain to the communications business, and offering Internet access is a key factor in this particular business area. The categories *Education & Research* and *Telephone & Communication* occupy the second and third positions with 5.32 % and 5.24 % respectively. Therefore, even today, companies in the area of Internet and information technology are still overrepresented because of their particular Internet affinity. A bit more unexpected, however, is that financial institutions seem well represented since the fourth position is taken by the *Finance & Insurance* cluster. All other categories are smaller with percentage values between 2.57 % (*Government & Military*) and 0.07 % (*Space*).

However, a limitation of our results is that the general *Company* cluster still encompasses 22.54 % of the classified ASs. This fact and the remaining number of unclassified ASs indicate that there is still a potential for improvement regarding the classification process. Yet it is questionable whether it is possible to reach much better results with semi-automatic classification approaches because of the presence of non self-explanatory organization names and acronyms such as NGM or EDP. Not only is it difficult to classify those simply based on keywords, it is also challenging to specify what kind of organization they represent without further manual and individual investigations.

5.2 SEC Classification

The industry-class frequency of organizations based on an alternative classification that is solely based on SEC data is shown in Figure 8. The industry classes *Construction & Manufacturing* as well as

Cluster	Definition	Result	Ratio
Address	ASs where no underlying organization is specified but an address, where the AS itself, the underlying organization or its managerial unit is located.	326	0.81 %
Company	Collecting bucket for those ASs which are hard to categorize based on their organization name but at least can be identified as a company.	9,104	22.54 %
Construction & Manufacturing	Mostly building firms and manufacturers are part of this class.	97	0.24 %
Consulting & Management	ASs related to advising and leading of a company.	139	0.34 %
Education & Research	ASs related to learning und gaining of new insights such as schools, universities, research facilities and networks as well as laboratories.	2,150	5.32 %
Entertainment & Information	ASs which are for example related to television, gaming, radio or publishing.	761	1.88 %
Finance & Insurance	This class consists mainly of banks and insurance firms.	1,664	4.12 %
Government & Military	ASs with an authority and military character as well as areal territories such as cities and states are relevant for this class.	1,039	2.57 %
Healthcare	Next to hospital (district) related entities, this class contains pharmaceutical firms.	695	1.72 %
IT & Internet Service	ASs that are affiliated with online as well as offline IT services and computer products. In general, this includes those firms which provide a service or product that is based on the Internet or IT, but which do not offer Internet access.	1,371	3.39 %
ISPs & Networks	Collects ASs of those organizations which offer Internet access or provide the necessary infrastructure.	3,775	9.35 %
IXPs	This class collects all ASs which function as exchange point in the Internet.	134	0.33 %
NIC	Contains those ASs which are "responsible for managing and allocating Internet resources" [6].	390	0.97 %
Space	Contains ASs of the area of astronautics.	30	0.07 %
Telephone & Communication	This class contains (mobile) telephone providers and sellers as well as general communication-based organizations.	2,118	5.24 %
Travel	Contains all ASs that are related to mobility and travel, such as airports, train stations, hotels and travel agencies.	103	0.26 %
Trade & Transport	Collects ASs of the area of wholesale and logistics including apparel and food.	232	0.57 %
Utilities	Organizations which provide electric power, water as well as other basic materials; also services such as waste disposal, coal and mining belong to this class.	256	0.63 %

Figure 7: Industry classes with their definitions and percentages based on the keyword classification.

Consulting & Management have the most organizations with ASs. The classes *Government & Military*, *ISPs & Networks*, *IT & Internet Service*, *IXPs*, *NIC*, as well as *Space* have no ASs at all. However, because not all companies are listed with the SEC and in particular governmental institutions and privately held companies are not registered, the lack of representation of these classes is inherent.

With the help of the SEC data it was possible to classify additional 116 ASs, which could not be classified via keywords only (Figure 9). Furthermore, the industry classes of 206 ASs could be specified more precisely which had previously been assigned to the *Company* class (Figure 10). Most newly specified classifications were assigned to the *Construction & Manufacturing* as well as the *Consulting & Management* industry classes. Their prevalence reflects the results of the SEC classification.

By combining both classification approaches we were therefore able to classify 22,892 of the 40,384 AS of the preprocessed list.

This accounts for 56.69 % of all considered ASs that could be assigned to an industry class. The final result is shown in Figure 11.

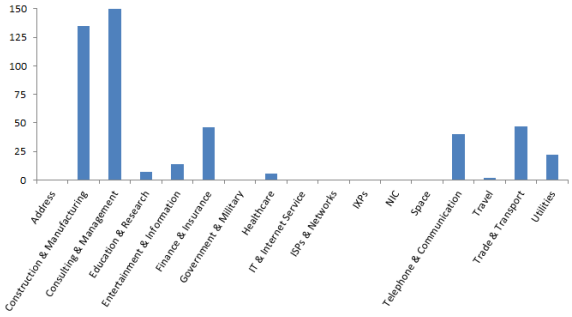


Figure 8: Industry class sizes based on SEC classification.

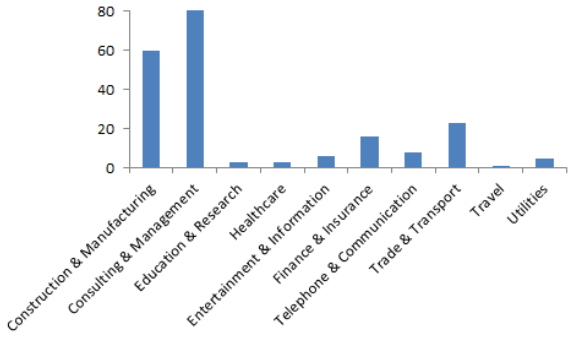


Figure 9: New classifications.

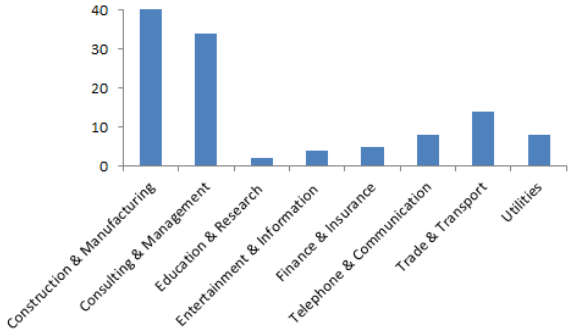


Figure 10: More precise SEC classification.

Cluster	Result	Ratio
Address	326	0.81%
Company	8,898	22.03%
Construction & Manufacturing	198	0.49%
Consulting & Management	254	0.63%
Education & Research	2,155	5.34%
Entertainment & Information	771	1.91%
Finance & Insurance	1,685	4.17%
Government & Military	1,039	2.57%
Healthcare	698	1.73%
IT & Internet Service	1,371	3.39%
ISPs & Networks	3,775	9.35%
IXPs	134	0.33%
NIC	390	0.97%
Space	30	0.07%
Telephone & Communication	2,134	5.28%
Travel	104	0.26%
Trade & Transport	2,546	6.30%
Utilities	269	0.67%

Figure 11: Final classification using both keyword and SEC data (Note: It was possible to categorize an AS into more than one industry class.).

6 CONCLUSIONS

This paper proposed a classification approach for categorizing ASs into detailed industry classes in order to better understand the economic background of the Internet structure. The industry classes are inspired by the NAICS (2013), which had the effect that an unprecedented level of detail regarding the industry classes for classification could be achieved. Data was mainly obtained from the CAIDA AS Rank project as well as from SEC.

The classification of ASs into industry classes based on their underlying organization revealed an on-going strong dominance of telecommunication and IT-related firms in the current Internet as well as of large institutions such as banks and universities.

It was possible to classify 56.69 % of all ASs (after preprocessing). Nevertheless, the amount of unclassified ASs indicates that there is room for improvement regarding the categorization process. A refined and extended keyword selection process could provide better results. Nevertheless, since there is a non-negligible amount of ASs having organizational specifications that are not self-explanatory or acronyms, this would involve a difficult challenge.

Some of our further explorative attempts to find new ways for AS classification with the help of clustering algorithms had limited success so far. However, another possible route could be to apply methods from Natural Language Processing (NLP) to the AS data and also for analysing search results from the Web for acronyms or other challenging organization names.

Moreover, customers of the various ISPs cannot be captured by the current method. It is often the case that large Internet providers also represent smaller customers who are not registered in the organizational information of the ASs. Here, studying the level of IP addresses could provide further insights but will also involve complex challenges.

Various other classification approaches might be feasible. In future work we will try to find other valuable classification systems aiming to take an even closer look at the composition of the Internet. Furthermore, we will use our classification results to further investigate the important players of the Internet backbone as well as to assess their economic interests and risks, at individual as well as global scales. Moreover, we aim to derive implications for Internet reliability and control assessments as well as for security and privacy analyses.

ACKNOWLEDGEMENTS

The authors thank Sebastian Dombrowski for his programming work during parts of this research.

REFERENCES

CAIDA, 2011. About CAIDA. http://www.caida.org/home/about/. (Access Dec, 2013).

CAIDA, 2012. AS Rank Project. http://as-rank.caida.org/?mode0=as-dump-info. (Access Dec, 2013).

CareerOneStop U.S. Department of Labor, Employment and Training Administration, 2013. NAICS-SIC Cross/Reference. http://www.acinet.org/industry/Ind_Sic.aspx?id=&nodeid=1. (Access Dec, 2013).

Chang, H., Jamin, S., Mao, Z., Willinger, W., 2005. An Empirical Approach to Modeling Inter-AS Traffic Matrices. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*.

Dhamdhere, A., Dovrolis, C., 2011. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Transactions on Networking* 19(5):1420–1433.

Dimitropoulos, X., Krioukov, D., Riley, G., claffy, k, 2005. Classifying the Types of Autonomous Systems in the Internet. *SIGCOMM 2005 Poster*, Philadelphia, Pennsylvania.

Dimitropoulos, X., Krioukov, D., Riley, G., claffy, k. 2006. Revealing the Autonomous System Taxonomy: The Machine Learning Approach. In *Passive and Active Network Measurement Workshop (PAM)*, Adelaide, Australia.

Freund, Y., Schapire, R.E., 1997. A Decision Theoretic Generalization of Online Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1):119–139.

International Telecommunication Union, 2011. Internet Users per 100 Inhabitants 2001-2011. http://www.itu.int/ITU-D/ict/statistics/material/excel/2011/Internet_users_01-11.xls. (Access Dec, 2013).

International Telecommunication Union, 2013. Key ICT Indicators for Developed and Developing Countries and the World. http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/ITU_Key_2005-2013_ICT_data.xls. (Access Dec, 2013).

NAICS, 2013. North American Industry Classification System. http://www.census.gov/eos/www/naics/. (Access Dec, 2013).

Potaroo, 2012. 32-bit AS Number Report. http://www.potaroo.net/tools/asn32/. (Access Dec, 2013).

RouteViews, 2013. University of Oregon Route Views Project. http://www.routeviews.org/. (Access Dec, 2013).

Réseaux IP Européens, 2011. AS Number Assignment Policies. http://www.ripe.net/ripe/docs/ripe-525. (Access Dec, 2013).

RIPE NCC, 2013. Réseaux IP Européens Network Coordination Centre. http://www.ripe.net/. (Access Dec, 2013).

U.S. Department of State, 2013. World Population Day 2013. http://www.state.gov/secretary/remarks/2013/07/211828.htm. (Access Dec, 2013).

United Nations Population Fund (UNFPA), 2001. The State of World Population 2001 – Demographic, Social and Economic Indicators. https://www.unfpa.org/swp/2001/english/indicators/indicators2.html. (Access Dec, 2013).

United States Securities and Exchange Commission (SEC), 2013. The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation. http://www.sec.gov/about/whatwedo.shtml. (Access Dec, 2013).

APPENDIX

Cluster	Keywords
Address	avenue, building, flat, floor, strasse, gpo box, handelsweg, mcpo box, no., po box, road, street, suite(s), tower
Company	associates, agency, a\.s\., bv, b\.v\., cjsc, co, co kg, companies, company, coporation, corp, corporation, d.o.o., de c.v., enterprise(s), gmbh, inc, incorporated, l\.l\.c, limited, llc, llp, lp, l\.p\., ltd(a), organization, s.a. de c.v., s.p.a., s.r.l., sp. z o.o., srl, s\.a\., s\.l\., sa, sas, sl, trust, z\.s\.p\.o, zspo
Construction & Manufacturing	architect(s), builders, building company, building society, construcoes, construction, constructora, constructors, electronics, machine, manufacturer(s), manufacturing, producers
Consulting & Management	beratung, business solutions, capgemini, consultance, consultancy, consultant(s), consulting, ernst & young, management company, pricewaterhousecoopers
Education & Research	.*universitaet, academic, academisch, colegio, college(s), desire2learn, ecole, education(al), fachhochschule, forschungsgemeinschaft, forschungsgesellschaft, fraunhofer, institute, instituto, knowledge network, laboratories, laboratory, labs, learning, mitre, physics, polytechnic, recherché, research, school(s), science(s), supercomputer, supercomputing, univ, universidad, universidade, universitaet, universitaria, universitas, universite, universiteit, universitesi, universitet, universiti, universities, university, univerzitet
Entertainment & Information	advertising, bbc, bertelsman, book(s), broadcasting, entertainment, football, fun, game, gaming, library, magazine(s), mcgraw-hill, media, marketing, medien, multimedia, news, newspaper(s), printing, publications, publishing, radio, reuters, television, times, tv, weather, zdf
Finance & Insurance	allianz, american express, asset management, assurance, banca, banco, bank, banka, banque, blue shield, capital, credit, finance, e*trade, financial, goldman, guggenheim, hsbc, insurance, investment, leasing, payment, real estate, reinsurance, rental, societe generale, stock exchange, stonepeak, visa
Government & Military	administration of, agency, air force, army, authority, board of, bureau of, city of(fice), committee, commonwealth of, congress(ional), council, county of(fice), department of, dept, district of(fice), dod, embassy, federated states, gov, government, house of, iles de, military, ministry, nato, navy, northrop grumman, parliament(ary), province of, senate, state of, united nations, united states postal service, US geological survey
Healthcare	bayer, blood, dental, drug(s), drugstore, elektromedizinische, emergency, health(care), hospital(s), johnson & Johnson, klinikum, medical, medicine, medizinische, merck, novartis, pfizer, pfizerswitzerland, pharma(cy), pharmaceuticals, pharmafarm, propharma, social security, transplant
IT & Internet Service	akamai, apple inc, computer hardware, computer products, computer science, computer service(s), computer software, computer solutions, computer systems, content provider, content service provider, content solution(s), data center(s), data corporation, data processing, data service(s), data solution(s), data systems, dell, fujitsu, general electric, google, hewlett-packard, host, hosting, ibm, information systems, information technology, internet service(s), internet systems consortium, it services, microsoft, neterra, network service(s), network systems, oracle, othello, samsung, sap, schuberg philis, siemens, sony, sungard availability, thinktech, verisign, web service(s), yahoo
ISPs & Networks	aol, arcor, at&t, backbone, broadband, bt italia, cable network(s), cogent, comcast, connection(s), esnet, exatel, fibernet, freenet, gts, iletisim hizmetleri, internet access, internet provider, internet service provider(s), internet solution(s), isp, lattelekom-apollo, level 3, linxtelecom, netassist, netcologne, network access, network provider, network service(s), network solution(s), networks, ntt america, optical network, prometey, qwest communication(s), reseau national, reseau regional, retn, road runner, rostelecom, singtel optus, smartcity, sprint, surfnet, swisscom, t-2, telecom, telekom, telia latvija, teo, time warner cable, towerstream, transit, true internet, uzbektelecom, verizon, versatel, vimpelcom, west call, wireless
IXPs	exchange point, internet exchange, internet exchange point, ix, ixp(s), link, open exchange, peering exchange
NIC	afnic, american registry, apnic, arin, east-ukrainian, internic, network information center, network infomation center, network information centre, nic, ripe ncc,
Telephone & Communication	alcatel, bell canada, communication(s), e-plus, elisa, ericsson, lambdarail, mobile, motorola, nokia, o2, phone, radiotelephone, rockefeller group, singtel optus, telecommunication(s), telecomunicaciones, telefonica, telekommunikation, telekomunikacije, telekomunikacja, telekomunikasi, telecomunicazioni, telephone(s), telianet, turkcell, vodafone
Trade & Transport	amazon, apparel, clothing, coca-cola, fedex, food(s), logistic(s), logisticare, retail(ers), shaya magazacilik, shipping, shoe(s), supply, trade, trading, transport(ation), wal-mart, wholsesale
Travel	air canada, airline(s), airport, bahn(hof), boing, flughafen, klm, lufthansa, hotel(s), railway, reisebuero, resort, travel, vacation
Space	aeronautics, aerospace, astronomy, nasa, space administration, space agency, space research, space telescope
Utilities	bp, coal, electric power, electricity, energy, farmer, farms, fiber, gas, mine, mining, offshore, petroleum, utilities, utility, waste, water

Exploring the Bitcoin Network

Annika Baumann, Benjamin Fabian and Matthias Lischke
Institute of Information Systems, Humboldt University Berlin, Spandauer Str. 1, 10178 Berlin, Germany
{annika.baumann, bfabian}@wiwi.hu-berlin.de

Keywords: Electronic Cash, Bitcoin, Network Analysis.

Abstract: This explorative paper focuses on descriptive statistics and network analysis of the Bitcoin transaction graph based on recent data using graph mining algorithms. The analysis is carried out on different aggregations and subgraphs of the network. One important result concerns the relationship of network usage and exchange rate, where a strong connection could be confirmed. Moreover, there are indicators that the Bitcoin system is a "small world" network and follows a scale-free degree distribution. Furthermore, an example of how important network entities could be deanonymized is presented. Our study can serve as a starting point in investigating anonymity and economic relationships in Bitcoin on a new structural level.

1 INTRODUCTION

Bitcoin is a decentralized digital currency based on a peer-to-peer network architecture and secured by cryptographic protocols. It was originally proposed by Nakamoto (2009). Anonymity and avoidance of double spending are realized via a block chain, a kind of transaction log that contains all transactions ever carried out in the network. In order to provide some anonymity, personal, identifiable information is omitted from the transaction. Therefore, the source and destination addresses are encoded in the form of public keys. Every public key, which serves as pseudonym, has a corresponding private key which is stored in an "electronic wallet". Private keys are used to sign or authenticate any transactions. To become part of the peer-to-peer network, one needs to install a client software that runs either on a local device or at cloud providers (Ober et al., 2013).

Authorization and verification are conducted by a complex proof-of-work procedure. Nakamoto (2009) proposed the use of a timestamp server which takes the hash of a block of items, timestamps it, and widely publishes the hash to the network. The proof-of-work also creates new Bitcoins in the network; this process is called "mining". Creation of Bitcoins is limited to a fixed amount of 21 million Bitcoins that can be introduced to the system; this limitation aims at avoiding inflation. Therefore, until that point is reached around the year 2140, money supply will increase at a certain rate every year (Drainville, 2012).

Our explorative work applies descriptive statistics and network analyses to the Bitcoin transaction graph. The network data was provided by Brugere (2013) who applied several tools for downloading and constructing the user network of Bitcoin. Several aggregations are used to highlight network characteristics. The research focuses on global time-varying dynamics within the network. As a first step of our methodology, qualitative research was conducted in order to gain an insight into related work and the transaction graph. Next, we explored the provided data and undertook required preprocessing steps for storing it appropriately in a database. Statistics and network analyses were conducted using this database; results were evaluated, interpreted, and compared to recent research on the transaction graph.

2 RELATED WORK

There are three main related research articles on the Bitcoin transaction graph that were published within the last two years. The most recent work carried out by Ober, Katzenbeisser and Hamacher (2013) focuses on time-varying dynamics of the network structure and the degree of anonymity. Using data of the period 03/01/2009 to 06/01/2013, the authors discovered that the entity sizes and the overall pattern of usage became more stationary in the last 12 to 18 months, which reduces the anonymity set.

The authors also show that the number of dormant coins is important to quantify anonymity. Inactive entities hold many of these dormant coins and thus further reduce the anonymity set (Ober et al., 2013).

Reid and Harrigan (2013) focus on anonymity in the Bitcoin network, analyzing the topology of the transaction and user network based on data of the time interval from 03/01/2009 to 12/07/2011. The authors adopt a preprocessing step to construct the user network. In order to improve the anonymity analysis, the researchers propose several methods including the integration of external information that is mainly held by businesses and other services which accept Bitcoin as payment. They show that it is possible to associate IP addresses from a public service with the recipient's public keys and link it to previous transactions.

In the third paper by Ron and Shamir (2013) the main focus lies on non-dynamic statistical properties of the transaction graph. The authors analyzed data of the period from 03/01/2009 to 13/05/2012, using various statistics such as distributions of addresses, incoming BTCs, balances of BTCs, number and size of transactions, and most active entities. They found that the majority of Bitcoins is not in circulation and that most of the transactions amount to a rather modest sum (less than 10 BTC). The researchers also analyzed the largest transactions in the network (greater than 50,000 BTCs) and determined their flows. They showed that most of these transactions are successors of the initial ones. Another interesting finding is that the transaction flows reveal some characteristic behaviors such as long chains, fork merge, and binary tree-like distributions (Ron, Shamir, 2013).

3 DATA MANAGEMENT

The data of the Bitcoin transaction graph is publicly available in order to enable the proof-of-work concept for verification of transactions. Sites such as *blockchain.info* or *blockexplorer.com* can be crawled for deriving the entire transaction graph. The data used by our work was collected and to some extent preprocessed by a project of the University of Illinois at Chicago (Brugere, 2013). It contains the time horizon from 01/03/2009 until 04/10/2013. We applied tools developed by Martin Harrigan and Gavin Andresen for extracting data from the *Bitcoin.dat* files in order to construct a user network according to the method introduced by (Reid and Harrigan, 2013). This procedure results in several raw text files (Brugere, 2013). The latest available data for download at the time of writing contained 230,686 blocks with around 37.4 million edges and 6.3 million nodes. The text files were transformed into a specific target format of two tab-separated files, one relationship file and one node file. Once the data had an appropriate structure, it was imported into a relational database. For analyzing the dynamics and topological characteristics of the graph structure, *NetworkX* was used (http://networkx.github.io/) (Hagberg et al., 2008).

4 ANALYSIS METHOD

In the first step of the analysis several descriptive statistics were calculated. Some of our results were earlier established by Katzenbeisser and Hamacher (2011) and at the Chaos Communication Congress in 2013. Characteristics such as user activity and transaction volume were linked to the Bitcoin exchange rate provided by *Mt.Gox*, which provides services for exchanging Bitcoins (https://www.mtgox.com/).

The second part of the analysis regards the network structure and topology. Since financial transaction networks are always evolving and not static, all measures were applied for different time horizons in order to investigate the dynamics. In the following the network measures are briefly introduced.

The *Degree* distribution captures the structure of networks in terms of the individual connectivity of nodes. The in-degree of a node i is the total number of connections to the node i and is the sum of the ith-column of the adjacency matrix. For the out-degree, the sum of the ith-row of the adjacency matrix is calculated (Gross and Yellen, 2004). One characteristic, often revealed by real networks, is that the degree follows a power law (Clegg, 2006), e.g., as shown by Barabasi, Albert and Jeong (2000) for the World Wide Web and by Inaoka, et al. (2004) in cases of financial transaction networks.

The *Average Clustering Coefficient* measures the global cliquishness on the graph. Watts and Strogatz (1998) applied the clustering coefficient in order to discover the small world phenomenon within several networks. The *Average Shortest Path Length* is defined as the average number of steps along the shortest paths for all possible pairs of nodes and measures the efficiency of information or mass transport in the network (Mao and Zhang, 2013). According to network theory one can determine how efficient Bitcoin is with respect to transactions.

Eigenvector Centrality measures the influence of

one node on other nodes. For each node it is defined as the value of the corresponding component of the principal eigenvector of the adjacency matrix defining the network. Accordingly, a node with a high eigenvector score is one that is adjacent to nodes that also have high eigenvector scores (Borgatti, 2005). This measure is essential for discovering central hubs such as exchanges, miners, or "laundry services" that are important nodes in the Bitcoin network.

5 ANALYSIS AND RESULTS

The descriptive statistics were applied over the entire time horizon from 01/03/2009 until 04/10/2013. The transaction value per day has a wide range beginning with the initial transaction of 50 BTC up to a daily amount of nearly 30 million BTC (19th September 2012).

Table 1: Statistics of the Bitcoin network.

Dataset:	03.01.2009 - 10.04.2013		*Transactions (Relations):* 37450461 *Economic Entities (Nodes):* 6336769				
	Median	*Mean*	*Sd*	*Skew*	*Min*	*Max*	*Correl [ExRate]*
Transaction Value [BTC]	173.457	910.053	2.231.647	7	50	29.958.714	0,199
Number of Users	1.637	4.049	5.243	2	1	36.120	0,730
Number of Transactions	3.678	24.084	38.303	2	1	189.284	0,680

The distribution of the transaction values is strongly skewed to the left. Another notable result is the high correlation between the number of active users, the number of transactions, and the MtGox exchange rate (BTC/USD), see Figure 1.

There are cutoffs at the beginning of trades when the dollar parity was achieved and maintained with negligible changes on 04/13/2011, and at the end when an extremely high exchange rate of around 237 BTC/USD was reached. Both figures show a high heteroscedasticity of the data. This indicates a highly speculative behavior in the network. The relationship will be investigated more thoroughly later on.

Table 2 shows the five largest entities in the network according to their number of public keys. The largest one has the entity ID 11 with over 318,221 public keys. One can also see that this entity is involved in the biggest transactions within the network. All the largest transactions are likely related to each other as indicated by the close time horizon and the entities involved.

Ron and Shamir (2013) conducted an analysis of these Bitcoin flows and came to the conclusion that nearly all major transactions are related. Another interesting result regards the huge amount of tiny transactions. The highest percentage of transactions (6.80%) according to their trade volume corresponds

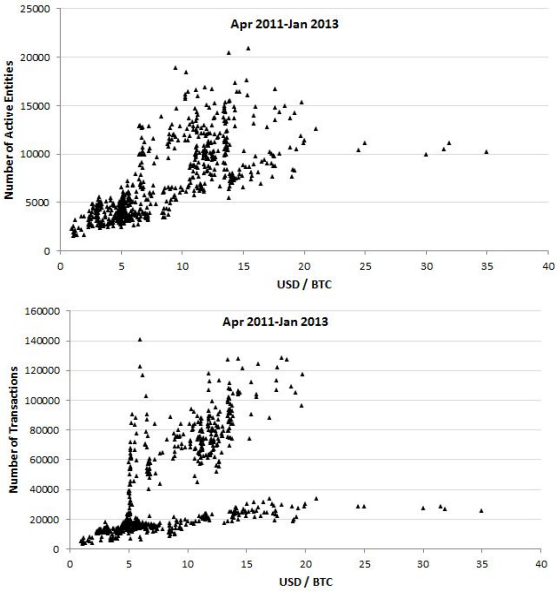


Figure 1: Correlations of user activity (left) and number of transactions (right) to exchange rate.

to the transactions of the range from 0.00000001 to 0.00001 BTC. Figure 2 shows the transaction values in a histogram, indicating the peaks of the highest transactions occurring.

Table 2: Transactions and users in the network.

Largest Entities		*Largest Transactions (Value)*				*Small Transactions (Value)*			
Entity ID	*# PubKeys*	*User_From*	*User_To*	*Date*	*Value*	*low*	*high*	*#*	*%*
11	318221	637193	637137	15. Nov 11	500000,00	0,00000001	0,00001	2546657	6,80%
29	209249	637137	11	16. Nov 11	499720,70	0,01	0,01099	2187115	5,84%
74	109128	11	636665	16. Nov 11	499643,98	0,0991	0,10009	777635	2,08%
12564	99939	636665	11	17. Nov 11	499609,08	...	...	...	...
27	64993	11	11	17. Nov 11	499420,95	49,6	50,59	344871	0,92%
						10	10,99	332412	0,89%

There is a strong relationship between the exchange rate of Bitcoin and the activity in the network. User activity increases immediately after a peak was reached by the exchange rate. A rolling window was constructed to investigate the relationship for different time windows. The user activity was measured for the last day, last 10 days, last 30 days and the last 100 days. Every rolling window shows a strong relationship but shrinks when extending the time horizon. The correlation coefficient for the last day is 0.736, last 10 days – 0.710, last 30 days – 0.671, and for the last 100 days – 0.641.

Figure 3 shows the BTC/USD exchange rate provided by the Bitcoin exchange Mt.Gox. There is a cutoff at the end of the time series due to a tremendous increase up to $237. In the following, some events are noted that might explain several of the strong movements in the exchange rate and the respective attention by more potential users of Bitcoin:

a) Start of the public trading of Bitcoins.
b) First time reaching dollar parity on 10th February 2011.
c) Several articles and media attention on Bitcoin, e.g., Forbes, Businessweek, or Bloomberg.
d) Abandonment of Paypal on Cyberlocker sites due to privacy concerns (Dotson, 2012).
e) Cyprus financial system about to collapse, Bitcoin is considered as new safe haven (see Mey, 2013).

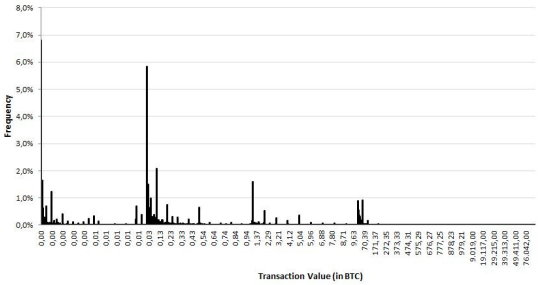


Figure 2: Histogram of the transaction value.

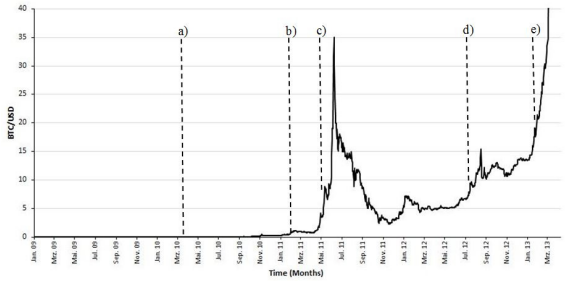


Figure 3: BTC/USD exchange rate and events.

Such a strong relationship can also be seen for the number of transactions carried out on the network. This is not surprising since higher activity of users leads to more transactions. The increasing number of transactions follows the exchange rate movements. The correlation coefficient to exchange rate is 0.680 and to user activity 0.928. Between transaction value and exchange rate there is a rather low correlation coefficient with 0.198.

In the following, the focus is placed on network structure. For this, the degree distribution was constructed. For every year since the start of Bitcoin in 2009, the degree k was calculated for every user entity by counting and summing ingoing and outgoing transactions (in- and out-degree). The resulting total degree distribution is drawn on a double logarithmic scale. The distribution also gives an insight into the network usage over time. In the beginning of network activity in 2009, there have been a lot of fluctuations. With increasing network usage the degree distribution seems to converge over time to a scale-free behavior that is also shown by many other real networks. In case of the Bitcoin network this means that the majority of users have a low degree while a small but non-negligible amount of users have a high degree k.

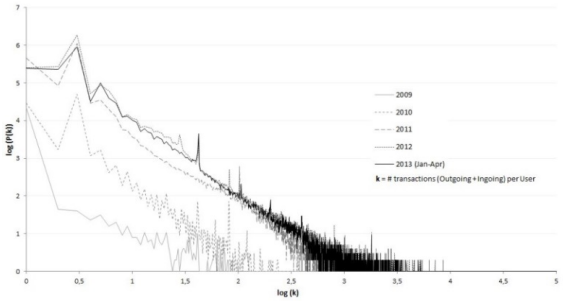


Figure 4: Degree distribution of the Bitcoin network.

Another important metric in the network analysis is the average clustering coefficient. In order to find evidence for a small world network, one can compare the Bitcoin graph to a random network with the same amount of nodes and edges like Watts and Strogatz (1998) did in their analysis. This measure was calculated on a monthly basis for the years 2012 and 2013. For 2011 the calculation was done quarterly. The years 2009 and 2010 were omitted from the analysis due to rather low activity in the network and lots of transactions between the same entities. Over time, the average clustering coefficient is rather high, indicating a small world network.

It can be seen that clustering decreases with increasing activity within the network. In quarter two and three of 2011 the lowest measure was calculated, while the user activity increased in that time period. The same effect can be noted for August 2012 and March 2013. Hence, higher user activity in the network reduces the global cliquishness in the graph.

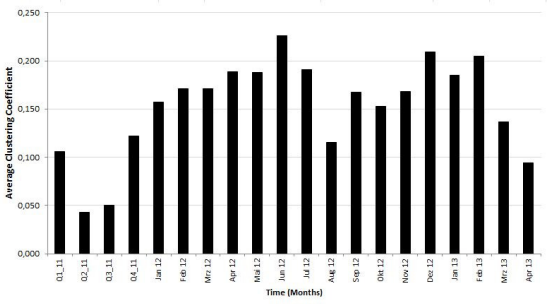


Figure 5: Average clustering coefficient over time.

Due to restrictions on computing power, the metrics average shortest path and the eigenvector centrality

are calculated just on the subgraph containing all transactions equal to and higher than 50,000 BTC. Since the average shortest path is only applicable on connected graphs, this metric is calculated for every connected subgraph within the network. In the large transaction network there are 11 disjoint subgraphs. The largest average path length is 125.083 and the lowest is 1.0. The first high value indicates a rather inefficient transfer of Bitcoins through the network according to a common interpretation of this measure. But since users are in control of transferring Bitcoins, this kind of inefficiency might be intended to obfuscate financial transactions.

The eigenvector centrality calculation did not converge to a solution within a reasonable time frame (on convergence see Hagberg et al. (2008)) thus only the degree centrality measure was used. The highest value occurs for the entity 11, which is also confirmed by the visualization of the largest hub in the graph. Degree centrality measures the importance of nodes within a network; the results show that the large transaction network node 11 is the most important node and can be considered as a hub for the others.

Table 3: Average shortest path length and degree centrality of the largest transaction graph.

Largest Transaction Sub Graph			
Sub Graphs	*Average Shortest Path Length*	*Entity ID*	*Degree Centrality*
1	125,083	11	0,0769
2	16,142	637193	0,0226
3	2,067	504303	0,00603
4	2,333	675451	0,00603
5	2,000	591334	0,00603
6	1,667	442450	0,00452
7	1,333	2132335	0,00452
8	1,333	...	...
9	1,000	636401	0,00301
10	1,000	...	...
11	1,000	5991405	0,0015

6 DEANONYMIZING ENTITIES

To demonstrate the possibility of deanonymizing at least some users in the Bitcoin network, the largest entity in terms of the number of public keys was selected. This entity 11 is also involved in the largest transactions that were carried out on the network. The first approach was to investigate the IP address belonging to the public key that initiates the transaction which is available from the site *blockchain.info*. It needs to be conceded that many IP addresses just reveal information (via Whois) about the last gateway before entering the block chain and thus cannot directly be associated with the real user. But one can receive information on the regional distribution of hosted services and their transactions. Users or business services accepting Bitcoins that are not using hosting services could be uncovered with this approach by using *getaddr.bitnodes.io*.

Another finding is that large and highly active entities providing exchange, laundry, mining, gambling services such as *Mt.Gox*, *SatoshiDice*, or *BTC Guild* are publicly known on the *blockchain.info*, and entity 11 could be identified as the exchange service Mt.Gox. To confirm this result, several transactions until April 10, 2013, in which Mt.Gox was involved were investigated using *blockchain.info* and could be linked to entity 11. Another method is to look up a particular public key of Mt.Gox in the dataset and show that it belongs to the public key pool of entity 11.

7 CONCLUSIONS AND OUTLOOK

Our research can serve as an exploratory starting point for the application of several techniques, descriptive statistics, and network analysis of the Bitcoin transaction graph. Recent results on the transaction graph were introduced. Standard descriptive statistics and more advanced methods in the field of network analysis were applied.

The results of the descriptive analysis show strongly skewed data series, especially for the transaction value. Another finding is the strong relationship between user activity, transaction volume, and the exchange rate of Bitcoin. One could also see that the largest entity is also involved in the largest transactions carried out in the network, and that the highest amount of transactions is of the smallest possible transaction size. Furthermore, the exchange rate was investigated and related to some events explaining its volatility. A strong relationship of user activity within different time horizons and the exchange rate could be demonstrated.

The network analysis revealed some new findings compared to previous research. We confirmed that the network degree distribution seems to converge to a scale-free network over time. A new contribution was the analysis of the average clustering coefficient, which is an indication for Bitcoin being a small world network as described by Watts and Strogatz (1998). The analysis of the average path length and degree centrality was conducted on a subgraph containing the largest transactions (>= 50,000) in the network. The results

show a very large average shortest path of around 125 for the largest connected subgraph, indicating inefficient user-driven transactions possibly aimed at hiding Bitcoin flows. Using the degree centrality measure, the largest hub in the subgraph, entity number 11 (also the largest entity in the entire network), could be found. Future work could aim for deanonymizing other major hubs of the network possibly by external information and experimental transactions.

The analyses conducted in this work could also be extended by further network measurements. One could investigate the small world character more thoroughly by advanced methods. Also further analysis on centrality can be conducted such as betweenness centrality or current flow betweenness centrality in order to get more insights on important hubs in the network. Further clique and clustering analysis can be used to expose social interaction characteristics of users.

One could also extend the data set with IP address and geo-location data in order to conduct novel research on the geographic characteristics of the network. Then it would be possible to analyze the network structure in different regions and how transactions occur between them. This can lead to a more thorough picture of structures and topology of the Bitcoin transaction graph. All of these analyses can also serve as a starting point in investigating anonymity and economic relationships in Bitcoin on a new structural level.

REFERENCES

Barabási, A-L; Albert, R; Jeong, H (2000): Scale-free Characteristics of Random Networks: the Topology of the World Wide Web. Physica A: Statistical Mechanics and its Applications 281(1–4): 69–77.

Borgatti, S (2005): Centrality and Network Flows. Social Networks 27(1): 55-71.

Brugere, I (2013): Bitcoin Transaction Network Extraction. https://github.com/ivan-brugere/Bitcoin-Transaction-Network-Extraction. (Access Dec 2013).

Clegg, R (2006): Power Laws in Networks. http://www.richardclegg.org/networks2/SpecialLecture_06.pdf. (Access Dec 2013).

Dotson, K (2012): Paypal's Abandonment of Major Cyberlockers May Become Bitcoin's Big Win. http://siliconangle.com/blog/2012/07/11/paypals-abandonment-of-major-cyberlockers-may-become-bitcoins-big-win/. (Access Dec 2013).

Drainville, D (2012): An Analysis of the Bitcoin Electronic Cash System. https://math.uwaterloo.ca/combinatorics-and-optimization/sites/ca.combinatorics-and-optimization/files/uploads/files/Drainville, Danielle.pdf. (Access Dec 2013).

Hagberg, A; Schult, D; Swart, P (2008): Exploring Network Structure, Dynamics, and Function Using NetworkX. In: Varoquaux, G; Vaught, T; Millman, J (Eds), Proceedings of the 7th Python in Science Conference: 11-15. Project documentation: http://networkx.lanl.gov/.

Inaoka, H; Ninomiya, T; Taniguchi, K; Shimizu, T; Takayasu, H (2004): Fractal Network Derived from Banking Transaction – An Analysis of Network Structures Formed by Financial Institutions. Bank of Japan Working Papers 04(04).

Katzenbeisser, S; Hamacher, K (2011): Bitcoin – An Analysis [28C3]. http://www.youtube.com/watch?v=-FaQNPCqG58. (Access Dec 2013).

Mao, G; Zhang, N (2013): Analysis of Average Shortest-Path Length of Scale-Free Network. Journal of Applied Mathematics, Vol. 2013, Article ID 865643.

Mey, S (2013): Bitcoin Rallye: Hackerwährung als sicherer Hafen für Sparer? http://blog.zdf.de/hyperland/2013/03/bitcoin-rallye-hackerwaehrung-als-sicherer-hafen-fuer-sparer/. (Access Dec 2013).

Nakamoto, S (2009): Bitcoin: A Peer-to-Peer Electronic Cash System. http://bitcoin.org/bitcoin.pdf. (Access Dec 2013).

Gross, J; Yellen, J (Eds) (2004): Handbook of Graph Theory, CRC Press.

Ober, M.; Katzenbeisser, S; Hamacher, K (2013): Structure and Anonymity of the Bitcoin Transaction Graph. Future Internet 5(2): 237-250.

Reid, F; Harrigan, M (2013): An Analysis of Anonymity in the Bitcoin System. In: Security and Privacy in Social Networks, Springer: 197-223. arXiv:1107.4524v2 [physics.soc-ph]

Ron, D; Shamir, A (2013): Quantitative Analysis of the Full Bitcoin Transaction Graph. Lecture Notes in Computer Science 7859: 6-24.

Watts, D; Strogatz, S (1998): Collective Dynamics of Small-World Networks, Nature, Vol. 393, June 1998: 440-442.

Organizational Maturity and Projects Performance
The Mediation of Benefits Management

Jorge Gomes[1], Mário Romão[2], Helena Carvalho[2] and Mário Caldeira[1]
[1]*Instituto Superior de Economia e Gestão, Universidade de Lisboa, Lisboa, Portugal*
[2]*ISCTE, Instituto Universitário de Lisboa, Lisboa, Portugal*
jorge.gomes@phd.iseg.utl.pt, mario.romao@iseg.ulisboa.pt, helena.carvalho@iscte.pt, caldeira@iseg.utl.pt

Keywords: Maturity Models, Project Performance, Benefits Management, IS/IT Investments.

Abstract: Nowadays the investments made by organizations to get better business performances are crucial and difficult to obtain. So, it´s more and more important that the business-cases underline not only the short-term objectives, but also the long-term benefits that promotes a sustainable future for organizations and investors. The main objective of our research is to study how investments on IS/IT influence projects performance in organizations. The rational of our model is that organizational maturity has a positive effect on projects performance with the mediation of benefits management. We emphasize that the integration between a maturity model and a benefits management approach can increase the effectiveness of strategic projects. Besides, it can also improve business confidence that the investments done match the desired maturity stages and will then collect benefits perceived as business value.

1 STAGE OF THE RESEARCH

The rational of our model is that organizational maturity has a positive effect on projects performance with the mediation of benefits management. We emphasize that the integration between a maturity model and a benefits management approach can increase the effectiveness of strategic projects. Besides, it can also improve business confidence that the investments done match the desired maturity stages and will then collect benefits perceived as business value. After a first year of intense literature review for our research project, we managed to deliver the research design conceptual model. At present, we are preparing the research design approach that includes the interview's guide and a survey to be launched in the beginning of 2014.

2 OUTLINE OF OBJECTIVES

This study is expected to help justify the rationale of IS/IT investments and reinforce the importance of internal organizational changes to achieve business benefits realization. The research questions that guide the present work are the following: (1) How do organizational capabilities enhance projects performance? (2) What tools can be used as a theoretical basis for analysing and implementing a methodology that enhances projects performance? (3) What are the critical factors that could cause projects to be successful?

3 RESEARCH PROBLEM

"Whether or not IS/IT spending does in fact lead to higher productivity?" This is an essential question that has been studied in the last 30 years by the academia. Brynjolfsson (1993) introduces the "mismanagement of information and technology" as one explanation for the productivity paradox. In line with his argument, other researchers argue that most organizations focus primarily on the implementation of technology rather than on the realization of the expected business benefits. Today's business environment can definitely be described as turbulent. Gaining a competitive advantage over competitors has been the focus of the organizations since a long time because only a sustainable competitive advantage can assure a long term existence in the market (Jugdev and Mathur, 2006); (Priem and Butler, 2001). Firms that have gained competitive advantage are attempting to maintain their

competitiveness by increasing knowledge and managing that knowledge (Porter, 1985). In a competitive environment, organizations need flexibility to meet customers' demands by offering customized and high-quality products and services. The notion that the IS/IT function serves different roles in different organizations and that such roles may evolve over time has been extensively discussed in past research. Although many studies have focused on the consequences of IS/IT investments (Brynjolfsson and Hitt, 1996); (Devaraj and Kohli, 2003), there have been fewer studies examining factors that impact the IS/IT capability. This research project will be based on the combination of the benefits management and maturity models to strengthen the final results of the IS/IT investments and enhance the expected benefits in order to improve projects performance. We claim, by linking maturity models and benefits management approaches, more effectiveness of the strategic projects and more improvement on the confidence of business sponsors that their investments will return the expected benefits. Benefits management adds value providing relevant information to the strategic projects and to enhance organizations to obtain higher levels of performance, identifying the goals and the benefits and clearly mapping the way to get them, supported on the right combination of organizational changes, enabling factors and IS/IT enablers (Ward and Daniel, 2006).

4 STATE OF THE ART

Some authors argue that the result of the studies that related investments in IS/IT and the increasing performance of the organizations in the last decades were far from true (Strassmann, 1997). However, others say that the amount spent in IS/IT and the business success has no direct connection (Hochstrasser and Griffiths, 1991). The relationship between IS/IT and productivity is widely discussed but little understood (Brynjolfsson, 1993). The idea that something is wrong with the investments in systems and information technology is not new, but nevertheless gain more visibility from the moment that Robert Solow (1987; 36), Nobel Prize in Economics, said ironically "we see computers everywhere except in the productivity statistics". This expression highlights the inability to demonstrate that investments in systems and information technology result in organizations productivity improvements, known as the "productivity paradox". Many organizations today need to deliver more complex products and services in a better, faster, and cheaper way. The business problems that companies address today require enterprise-wide solutions that call for an integrated approach and effective management of organizational resources to achieve business objectives. A maturity model approach is a process-driven improvement that provides organizations with the essential elements for effective change. It can be used to guide process improvement across a project, a division, or an entire organization. Benefits management reinforces the distinction between project results and business benefits. The main focus of investment lies not only in technology implementation, but mainly in improvements in organizational performance and business efficiency, e.g. improving processes and changing the ways the work is performed.

4.1 Maturity Models

For organizations to succeed in the global business competition of today, it is necessary that they produce a high standard of performance. Basically, the purpose of the maturity models is to provide a framework for improving an organization's business result by assessing the organization's strengths and weaknesses, enabling comparisons with similar organizations, and a measure of the correlation between organizations (Combe, 1998); (Ibbs and Kwak, 2000). In the IS/IT discipline, maturity is regarded as "a measure to evaluate the capabilities of an organization" (Rosemann & De Bruin, 2005:1). Levin and Skulmoski (2000) point out that the maturity models provides a framework to help enable organizations to increase their capability to deliver projects on schedule, within budget and according to the desired technical specifications. Working with different types of projects within an organization requires standard models in order to deliver successful future projects repeatedly, improve both the quality of future projects and gain knowledge and learn from past mistakes. Measuring maturity in organizations is regarded as a subjective instead of an objective measurement, since most significant research is primarily focusing on what people are doing operationally (Andersen and Jessen, 2003). Skulmoski (2001) recommend a view where competence and maturity should be linked together for project success and not focusing only on action and where competence should be regarded as a combination of knowledge, skills and attitudes that supports performance. The assessment procedures helped an organization understand where they have

been, where they are, and what processes they need to implement, in order to continue their implementation of management methodologies. The work of Ibbs and Kwak (1997, 2000) and Ibbs and Reginato (2002) focused on recognizing the benefits of investment in project management competency through measures of maturity in an organization's practice of project management. The assessment of maturity typically involves variations along five developmental stages (Jugdev and Thomas, 2002): Level 1: Initial (ad hoc) - Processes are not usually documented. Level 2: Repeatable or structured (abbreviated, planned) – Basic management practices have been established. Level 3: Defined (organized, institutionalized) - The management and technical processes necessary to achieve the organizational purpose will be documented, standardized and integrated to some extent with other business processes. Level 4: Managed (integrated) - There will be evidence of quantitative objectives for quality and process performance, and these will be used as criteria in managing processes. Level 5: Optimized (adaptive, sustained) - The organization will focus on optimization of its quantitatively managed processes to take into account changing business needs and external factors. The underlying assumption in the maturity models is that there is a relationship between higher levels of maturity and improved organizational Performance Projects.

4.2 Benefits Management

As competition increases as a result of globalization and other market factors, it is even more important that an organization performs at its best capabilities (Ashurst and Doherty, 2003). The decision making process over IS/IT investments is not as objective and transparent as it is claimed to be, creating significant failures on the benefits achievement process (Berghout et al., 2005). Since 1995, due to the recognition of the importance of benefits realization and management within different sectors, various approaches and models have been developed to help organizations identify, monitor and ultimately achieve the benefits. It is unlikely that benefits will simply emerge, as if by magic, from the introduction of a new technology. Their realization needs to be carefully planned and managed (Lin and Pervan, 2003); (Markus, 2004). Benefits are often identified in the early stages to form the business case and to sell the idea to the stakeholder's. A follow-up procedure with the purpose of evaluating those benefits achievement is often missing, and problems arise after the system delivery, when it's time to show if those previous stated benefits have been realized (Remenyi et al., 2007). The perception of the continuous unsuccessful IS/IT investments found a new way and approach for how projects are undertaken. The focus should be on the realization of the benefits, since that is the organization main reason to the investment (Ward and Daniel, 2006). A common characteristic of many unsuccessful programs is the vagueness with which the expected benefits are defined (Reiss et al., 2006). Without clearly defined objectives it is difficult to maintain focus when subsequent problems occur. The increased interest in benefits realization has coincided with the increasing use and complexity of IS/IT (Ashurst and Doherty, 2003); (Bradley, 2006). A benefit is an outcome whose nature and value are considered advantageous by an organization (OGC, 2007). Bradley (2006) defines it as an outcome of change which is perceived as positive by a stakeholder. The important point in the above definitions is that advantage is owned by individuals or groups who want to obtain value from an investment (Ward and Daniel, 2006). The benefits to an organization from IT- enabled change essentially emerge from three main reasons: Either stopping doing activities, doing better what have always being done, or even doing completely new things (Peppard and Ward, 2005). Ward and Daniel (2006) differentiate benefits as tangible and intangible, whether intangible benefits are those that can only be judged subjectively and tend to employ qualitative measures. In our research we follow the Cranfield Management School approach (Ward and Daniel, 2006) that includes a process model for benefits management (Figure 1):

(1) Identifying and structuring benefits - The proposed benefits are identified and, for each proposed benefit, business measures are developed, both financial and non-financial.

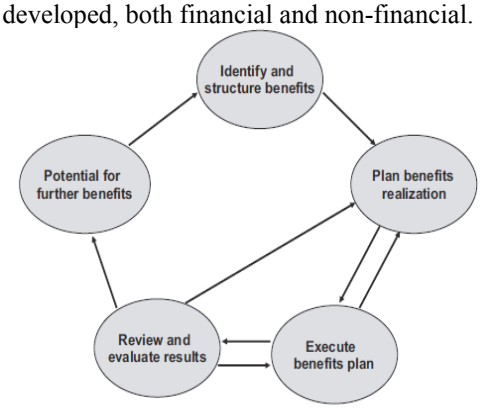


Figure 1: Process model for benefits management (Ward and Daniel, 2006; 105).

(2) Planning benefits realization - For each benefit, specific responsibility for realizing the benefit is allocated within the business. (3) Executing the benefits realization plan - Alongside the implementation of the proposed IT application, the necessary business changes as detailed in the benefits realization plan are carried out. (4) Evaluating and reviewing results – Following the full implementation of IS/IT and business changes, the previously developed business measures are used to evaluate the effects of the project. (5) Establish the potential for further benefits – As a result of the post-project review; it may become apparent that further benefits are now achievable.

Several authors recognized the importance of the two initial phases. Bennington and Baccarini (2004) suggest that the benefits identification should be a combined approach of interviews and workshops involving key stakeholders. Remenyi et al., (2000) note that a key aspect of benefits identification process is that the stakeholders learn to better understands what is required from an investment and what is affordable and possible. Best practice is to involve key stakeholders to identify and agree desired benefits maximizing the likelihood of commitment to realize those benefits across a range of levels in the business or the organization (Ward and Daniel, 2006). The key tool of Cranfield School approach is the benefits dependency network (BDN) (Figure 2) designed to enable the investment objectives and their resulting benefits to be linked in a structured way to the business, organization and IS/IT changes required to realize those benefits.

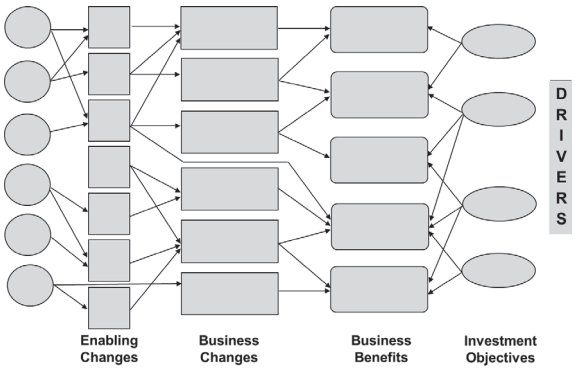


Figure 2: BDN framework (Ward and Daniel, 2006; 134).

Once the BDN has been constructed, measures for each benefit, responsibilities for achieving all the benefits and implementing all the changes, and timeframes must be established. The organization have setting targets, identifying and quantifying the benefits removing known problems through new IS/IT means and new ways of executing business processes and activities.

5 CONCEPTUAL MODEL

Based on the previous comments we proposed the research conceptual model shown in Figure 3

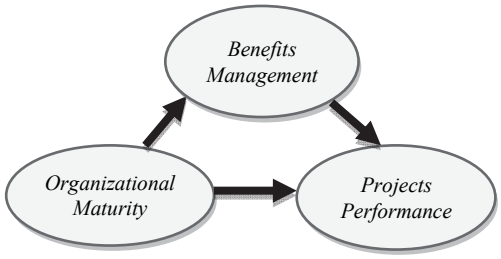


Figure 3: Mediation model.

In this model, organizational maturity is a construct consisting of seven dimensions represented by the following first order factors (Figure 4).



Figure 4: Organizational maturity construct (OGC, 2010).

Similarly, benefits management is conceptualized as a construct consisting of five dimensions, represented by the first order factors depicted in Figure 5. The rational of our model is that organizational maturity has a positive effect on projects performance with the mediation of benefits management. Organizational maturity will be assessed by using the P3M3 self-assessment questionnaire for practitioners. Concerning projects performance, the questions will be about how well the projects were performed during the last three years when comparing with all direct competitors in terms of financial results, market share, and sales

growth.

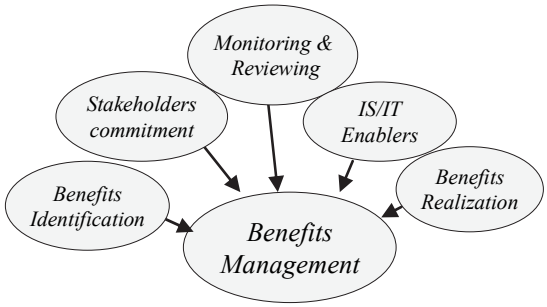


Figure 5: Benefits management construct.

Accordingly, we develop the following hypothesis:

(H1) – The relationship between organizational maturity and projects performance is mediated by benefits management. In order to test the mediation model, Baron and Kenny steps (1986) will be followed: (1) There is a significant relationship between organizational maturity and benefits management. (2) There is a significant relationship between organizational maturity and projects performance. (3) There is a significant relationship between benefits management and projects performance. If the effect of organizational maturity in projects performance remains statistically significant, even in the presence of the effect of benefits management (the mediator variable), then a partial mediation will be found.

6 METHODOLOGY

According to Silverman (2006) methodology is a general approach that establishes the ways in which any phenomenon can be studied. Yin (2003; 20) defines a research design as the "logical plan for getting from here to there, where here are the initial set of research questions to be answered, and there are the conclusions". As the scientific rigor in this kind of research paradigm is of crucial importance in order to truthfully explain the social reality under investigation, the study defined and adopted a research model and a set of hypotheses from the existing literature and sound theories. The set of hypotheses will then be empirically tested and validated under a mediation model. This investigation is primarily quantitative with its roots on the research model hypotheses testing. Furthermore, the research also draws on a preliminary qualitative component with the objective of defining and validating the constructs and its components through experts' interviews. The interviews are conducted in a semi-structured manner, in order to explore the usage of maturity models and benefits management concepts. The data collected through those interviews is then analysed and coded. The results of this exploratory study will allow us to validate the central constructs of our model and will serve as an input to the next step of the study: the validation of the research. The hypotheses testing will enable us to understand the nature of certain relationships amongst the variables of interest for the study. This will be based on a cross-sectional survey since it excludes explicitly the time dimension. Therefore, our research uses a mix of approaches, qualitative and quantitative, to address relatively complex related issues, not always perceived and applied in the same way in organizations.

7 EXPECTED OUTCOME

Although benefits management is emerging as one of the approaches that assist organizations to manage the whole life cycle of programmes and projects, there appears to be no evidence in the literature of the successful implementation of any of the approaches available. We hope to create clear evidence that benefits management contributes to projects success, namely: (1) In the identification of objectives and benefits of IS/IT investments in order to reinforce the organizational changes. (2) In the engagement of different stakeholders to define the expected benefits. (3) To create greater awareness of how projects outputs would lead to the achievement of outcomes by establishing a specific workforce to help defining and measuring the achievements. (4) To establish a formal method to plan and evaluate the expected benefits of investments, complying with initial requirements. (5) To create an environment for learning and improving, driven by expected benefits, as this was not the focus of the implementation. Investments in IS/IT projects and other change programmes are the means of improving projects and organizational performance, creating new strategic options and capabilities. By linking maturity models and projects performance, having benefits management as a mediator factor, we will try to prove that is easier to implement best practices on the organizations that have higher levels of maturity. We also want to confirm that a correct combination of IS/IT implementations and management practices lead to a positive influence on the projects performance. Then we aim to conclude

that organizations with higher levels of maturity have a positive influence on their projects performance, particularly in the cases where they make use of benefits management practices.

REFERENCES

Andersen, E. S., & Jessen, S. A., 2003. Project maturity in organizations. *International Journal of Project Management,* 21, 457-461.

Ashurst, C., & Doherty, N. F., 2003. Towards the formulation of "a best practice" framework for benefits realization in IT projects. *Electronic Journal of Information Systems Evaluation*, 6, 1-10.

Baron, R. M., & Kenny, D. A., 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology,* 51, 1173-1182.

Bennington, P. & Baccarini, D., 2004. Project in IT projects – An Australian perspective. *Project Management Journal*, June.

Berghout, E., Nijlan, M., & Grant, K., 2005. Seven Ways to get Your Favored IT Project Accepted – Politics in IT Evaluation. *Electronic Journal Of Information Systems Evaluation,* 8, 31-40.

Bradley, G., 2006. *Benefit Realization Management: A Practical Guide for Achieving Benefits through Change*. Gower Publishing.

Brynjolfsson, E., & Hitt, L., 1996. Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending. *Management Science*, 42, 4, 541-558.

Brynjolfsson, E., 1993. The Productivity Paradox of Information Technology. *Communications of the ACM,* 36, 12, 67-77.

Combe, M. W., 1998. Standards Committee Tackles Project Management Maturity Models. *PM network*, 12, 8, December.

Devaraj, S., & Kohli, R., 2003. Performance Impacts of Information Technology: Is Actual Usage the Missing Link? *Management Science*, 49, 3, 273-289.

Hochstrasser, B. & Griffths, C., 1991. *Controlling IT Investments: Strategy and Management*. Chapman & Hall, London.

Ibbs, W. C., & Reginato, J., 2002. Can good project management actually cost less? *Proceedings of the 33rd Annual Project Management Institute*. Seminars and Symposium, San Antonio.

Ibbs, W. C., & Kwak, Y. H., 2000. Assessing project management maturity. *Project Management Journal*, 31, 1, pp. 32-43.

Ibbs, W. C., & Kwak, Y. H., 1997. *The benefits of project management: Financial and organizational rewards to corporations.* Newtown Square, PA: PMI Educational Foundation.

Jugdev, K., & Mathur, G., 2006. A Factor Analysis of Tangible and Intangible Project Management Assets. *4th Project Management Research Conference.* Montreal, QC, PMI.

Jugdev, K., & Thomas, J., 2002. Project Management Maturity Models: The silver bullet of competitive advantage? *Project Management Journal*, 33, 4-14.

Levin, G., & Skulmoski, G., 2000. The project management maturity. *ESI Horizons*, 2, 3, 1-7.

Lin, C., & Pervan, G., 2003. The practice of IS/IT benefits management in large Australian organizations. *Information & Management*. 41, 31-44.

Markus, M. L., 2004. Techno change management: using IT to drive organizational change. *Journal of Information Technology,* 19, 4-20.

OGC, 2010. *P3M3 - Project Management Self-Assessment*. Office of Government Commerce.

OGC, 2007. *Managing Successful Programmes.* Office of Government Commerce.

Peppard, J., & Ward, J., 2005. Unlocking Sustained Business Value from IT Investments. *California Management Review,* 48, 1, 52-70.

Porter, M., 1985. *Competitive Advantage: Creating and Sustaining Superior Performance*. New York: Free Press.

Priem, L., & Butler, E., 2001. Is the resource-based view a useful perspective for strategic management research? *Academy of Management Review*, 26, 1, 22-40.

Reiss, G., Anthony, M., Chapman, J., Leigh, G., Pyne, A., & Rayner, P., 2006. Gower Handbook of programme management. *Gower Publishing.*

Remenyi, D., Money, A., & Bannister, F., 2007. The Effective Measurement and Management of ICT Costs and Benefits. Burlington, *MA: CIMA Publishing*.

Rosemann, M., & de Bruin, T., 2005. Towards a business process management maturity model. *ECIS2005 - 13th European conference on information systems,* Regensburg.

Silverman, D., 1998. *Harvey Sacks: Social science and conversation analysis.* Cambridge: Polity.

Skulmoski, G., 2001. Project Maturity and Competence Interface. *Cost Engineering*, 43, 11-19.

Solow, R., 1987. *We'd better watch out*. New York Times Book Review, July 12, 36.

Strassmann, P., 1997. The Squandered Computer. *The Information Economics Press*, New Canaan, Connecticut.

Ward J. & Daniel, E., 2006. *Benefits Management, Delivering Value from IS and IT Investments.* Chichester, UK: John Wiley & Sons.

Yin, R. K., 2003. *Case Study Research, Design and Methods.* 3rd ed. Thousand Oaks: Sage Publication

Technology-enabled Bullying & Adolescent Non-reporting
Breaking the Silence

Justin Connolly[1] and Regina Connolly[2]
[1]*Dublin City University, Dublin 9, Dublin, Ireland*
[2]*Dublin City University, Business School, Dublin 9, Dublin, Ireland*

Keywords: Information and Communication Technologies, Technology-enabled Bullying, Cyberbullying.

Abstract: Although early research has pointed to the fact that the successful intervention and resolution of cyberbullying incidents is to a large degree dependent on such incidents being reported to an adult caregiver, the literature consistently shows that adolescents who have been bullied tend not to inform others of their experiences. However, the reasons underlying reluctance to seek adult intervention remain undetermined. Understanding the factors that influence adolescent resistance will assist caregivers, teachers and those involved in the formulation of school anti-bullying policies in their attempts to counter the cyberbullying phenomenonre should be a space before of 12-point and after of 30-point.

1 INTRODUCTION

The problem of adolescent bullying has evolved in tandem with the digitization of society. Bullying is a problem that transcends social boundaries and can result in devastating psychological and emotional trauma including low self-esteem, poor academic performance, depression, and, in some cases, violence and suicide. In its traditional context, it has been described as being characterized by the following three criteria: (1) It is aggressive behavior or intentional 'harm doing' (2) which is carried out repeatedly and over time (3) in an interpersonal relationship characterized by an imbalance of power. One might add that the bullying behavior often occurs without apparent provocation," and "negative actions can be carried out by physical contact, by words, or in other ways, such as making faces or mean gestures, and intentional exclusion from a group (Olweus 1999, pp.10-11).

Cyberbullying, which is bullying conducted through the medium of electronic communication tools (such as email, mobile phone, social networking sites, Personal Digital Assistants, instant messaging tools and the World Wide Web) has been defined by Willard (2007) as '… being cruel to others by sending or posting harmful material or engaging in other forms of social cruelty using the Internet or other digital technologies, such as cell phones. Young people may be the target of cyberbullying from others or may engage in such harmful behavior. Direct cyberbullying involves repeatedly sending offensive messages. More indirect forms of cyberbullying include disseminating denigrating materials or sensitive personal information or impersonating someone to cause harm (p.10).'

As can be seen from the above definitions, there are important distinctions between cyberbullying and bullying. The first distinction relates to the nature of the bullying. Traditional forms of bullying are usually direct and bullies are visible, while cyberbullying can be anonymous and bullies in cyberspace do not have to be physically stronger or bigger than cyber-victims. Second, bullying occurs often at a particular time and place, whereas cyberbullying can happen anytime, anywhere including in the home. Third, cyberbullying can spread exponentially faster (e.g. copy and paste a message and send it around the world) than traditional forms of bullying. Fourth, cyberbullying can be preserved easily (such as saving messages on a phone, memory stick, disk, etc). Fifth, bullies usually have poor relationships with teachers but it has been noted that cyberbullies can have good relationship with teachers (Ybarra & Mitchell, 2004). Finally, on the one hand, bullying commonly happens on school property; cyberbullying, on the other hand, frequently occurs outside school property which makes identifying and combating such behaviour much more difficult.

2 NON-REPORTING

Although even early research in this field (e.g. Olweus, 1993) has pointed to the fact that the successful intervention and resolution of bullying incidents is to a large degree dependent on such incidents being reported to an adult caregiver, the extant literature consistently shows that adolescents who have been bullied tend not to inform others of their experiences (e.g. Petrosino et al., 2010; Black, Weinles and Washington, 2010; Mishna and Alaggio, 2005; Naylor and Cowie, 1999; Charach, Pepler and Ziegler, 1995). For example, Ybarra et al (2006) found that 33% of victims of cyberbullying in their study did not tell anyone about the incident. More recent work by Petrosino et al. (2010) is of particular importance given that their data collection involved a nationally representative sample from American schools and was based on figures from the National Crime Victimization Survey School Crime Supplement (2007), which showed that 64% of adolescents between the age of 12 and 18 did not report their experience.

Smith, et al.'s (2006) study of 92 students across the UK found that almost one-third of students who acknowledged being targetted by cyberbullies chose not to speak about their experience when it happened. This figure is close to the findings of the NCH (2005) study, which revealed that 28% of those targetted by cyberbullies chose to remain silent rather than seek help in resolving the problem. Slonje and Smith (2008) state: 'Cybervictims most often chose to either tell their friends or no one at all about the cyberbullying, so adults may not be aware of cyberbullying' (p.147).

Compounding the problem is the fact that variation appears to exist between cultures. For example, a cross-cultural comparative study by Li (2008) showed that 9% of Canadian students reported their cyberbully experience to adults and less than one-fifth of those aware of a cyberbullying incident reported the issue to an adult. In comparison, 66% of Chinese students who experienced cyberbullying informed an adult and 60% of 'bystanders' reported the phenomenon to an adult (p.7). The reason for this significant difference in reporting behaviours between both countries and cultures may result from a combination of sociological and philosophical reasons deeply ingrained in the respective cultures.

3 RESEARCH INSIGHTS

Despite the fact that many studies have found that adolescents do not report their bullying or cyberbullying experiences, there is a dearth of empirical work examining the reasons for same. Two studies deserve particular mention. The first is a quantitative study by Holfeld and Grabe (2012), which replicated earlier descriptive research on the prevalence of cyberbullying and examined why students do not report cyberbullying. Using a sample of 383 students from four middle schools in a North American city (with average student age of 13.5 years), and using a subset of self developed measures to capture non-reporting (4 questions in relation to own experience and 3 questions in relation to reporting of peer experiences) , they found that 16% of students reported being cyberbullied in the previous year and of those 62% were cyberbullied at least once or twice in the last 30 days. Only 11% of students reported cyberbullying others at least once in their lifetime and 9% in the last year. Cell phone cyberbullying was the method used most frequently. As a key point of that study concerns the reporting aspects of cyberbullying, the findings show that almost 30% of students who were cyberbullied in the past year did not report the incident. When asked to explain their reporting behaviour, 57% of the responses comprised they didn't feel it was a big deal or they felt they could handle it on their own. 29% of students considered that reporting would make it worse or were scared to tell. Whilst this study provides an empirical attempt to understand the issue of non-reporting in more depth, Holfeld and Grabe's work is limited in the sense that the number of questions used to capture non-reporting comprises a small number of self developed measures that are not validated or tested for reliability and the study was purely quantitative in nature and not followed up by in depth exploration of the issue. It is likely that our understanding of the factors that influence adolescent non-reporting would benefit from a triangulated approach to data collection.

A second study that has sought to bring greater clarity to this issue is that of De Lara (2012). Using a qualitative approach, she studied the non-reporting problem in four schools (two rural and two urban) in the New York region. The sample comprised twelve focus groups (3 in each school, comprising 97 students) and 51 individual interviews (with some cross-participation between students being involved in both focus group and individual interview) of which 52% were female and 48% male adolescents.

A significant finding of the research is that the reasons for non-reporting appear to be multifactorial with the results indicating that the adolescents in this sample did not report their experiences due to the ubiquitous nature of bullying; a sense of helplessness; concerns over inappropriate adult action; self-reliance; shame; parental omniscience; and a different definition of bullying than adults use (De Lara, 2012: 288).

Interestingly, students in the research considered bullying to be the norm or something to be expected whilst witnesses to such behaviour also perceived it as a normal rite of passage in school. They were despondent about the potential for successful adult intervention, as they feared that parental intervention could make things worse or, at the other extreme, that adults would not take the concern seriously enough. It was of particular concern that some reported being told by teachers to deal with the problem themselves – an obvious flaw in the duty of care by individual teaching staff. When adolescents seek help from an adult and the bullying continues unabated despite reporting the issue, the research shows that they are likely to withdraw from communicating the issue further to the adult caregiver (DeLara, 2008; Garbarino & DeLara, 2002). This confirms the findings of Petrosino et al. (2010) and Pepler et al. (2008) whose research shows that between 40-65% of adolescents never report their experience of bullying to an adult.

An indirect finding of DeLara's work that may also provide insight as to the reasons for non-reporting relates to the interpretive difference that adolescents and adults attribute to the term "bullying". For example, she found that many of the students in her sample, when asked to define bullying, described it as: 'when someone is *mean* to me' [italics mine]. However, 'mean' behaviour – is not reflected in current descriptions of bullying in the literature. This difference in interpretation was previously highlighted by Smith, Cowie, Olaffson, and Liefooge (2002) who found that adolescent perception of what constitutes bullying could differ remarkably from that of adults. For example, a student experiencing sexual harassment may not be aware that they are being bullied. It follows, therefore, that if students' understanding of bullying differs from adult understanding, the chances of reporting and intervening in such bullying behavior are reduced. This is of significance in implementing anti-bullying policies, educating students as to what constitutes 'bullying' and thereby encouraging them to come forward to relate their experiences to a significant caregiver.

Whilst the work of De Lara (2012) is valuable in that it represents an attempt to examine the issue of non-reporting, it was limited to four schools (2 urban and 2 rural) within the New York region, the sample was small and therefore the generalisability of its findings remain uncertain. It is possible that local and cultural factors may have impacted the reasons for non-reporting behaviour. However, whether this is the case can only be determined by additional research on this issue in other and broader contexts. Further study in an Irish context would be particularly valuable not only in providing insight into this issue in relation to the factors that predict Irish adolescent non-reporting, but also as a comparative measure to establish the culture independence of these factors.

4 THE IRISH CONTEXT

Despite awareness of a considerable number of adolescent deaths in Ireland that have been related to cyberbullying (most notably Lara Burns, Erin Gallagher, Ciara Pugsley and Leanne Wolfe), empirical research on this issue remains remarkably limited. The death by suicide in America of another Irish adolescent, Phoebe Prince, brought worldwide attention to the gravity of the problem but the factors linking cyberbullying and death by suicide need deeper research.

Whilst there is a dearth of empirical research on cyberbullying in Ireland, concern about this issue is widespread and would appear to be justified. For example, a joint Irish Independent and 'Prime Time Investigates' survey (Irish Independent, 2008) of students found that approximately 30% of students have endured all types of bullying at secondary schools within a 3-month period of 2008. It also found that 1 in 5 schoolgirls had experienced cyberbullying as compared to 1 in 8 boys. The research data provided by the Anti-Bullying Research and Resource Centre at Trinity College Dublin in 2008 revealed an unsettling picture of the growth in online and mobile phone intimidation among secondary school pupils and showed that children as young as 12 are being targeted through mobile phone calls, text messages, e-mails, internet forums, chat rooms and social networking sites. Recent research on the prevalence and nature of cyberbullying was conducted by Cotter and McGilloway (2011) and comprised a sample of 122 adolescents from two secondary schools in the South of Ireland. The findings showed that although cyberbullying within that sample appeared to be less

prevalent than traditional bullying, the adolescents concerned considered it to be worse than traditional bullying, with the exception of email.

Whether individual factors such as age, gender, ethnicity, family cohesion (unity) or religious belief influences cyberbullying remains undetermined. Equally, the influence of situational factors such as attendance at public or private school on cyberbullying outcomes has not received adequate attention. Understanding the influence of these variables would contribute to parents and educators' insight into the problem and increase their ability to address it. (O'Higgins-Norman and Connolly, 2011).

In order to understand the hesitance that predicts adolescent non-reporting of bullying experiences, it is important that the views of those who experience and witness such behaviour should be understood and factored into any intervention or policy formulation processes. However as school policies are directed from government level and implemented by individual school boards which are constituted by adults and therefore such policies are designed and implemented from an adult perspective. Therefore, in order to effectively address the problem of bullying, it is imperative that research on this issue should consider the views of the adolescents who actually experience the behaviour. As DeLara (2012) states:

The preponderance of research on bullying tends to neither include the perceptions of students nor provide understanding about their reluctance to rely on adults for intervention. Research has found that students may not tell adults about bullying experienced or witnessed despite repeated encouragement and directives from adults (p.288).

4.1 Irish Non-reporting Behaviour

In Ireland, the issue of non-reporting of cyberbullying behaviour was initially identified by O' Moore & Minton (2011) who found that a distinct contradiction exists between intent and actual practice in terms of Irish adolescents reporting their cyberbullying experiences to adults. For example, they reported that whilst 14.6% of pupils stated that they *would* inform an adult at school if they were cyberbullied, *in reality*, only 6% of these pupils had actually reported their cyberbullying experience. Instead, the found that pupils were over twice as likely to do nothing at all, five times more likely to send an angry message back, and five times more likely to talk to a friend.

Recent research by Cotter and McGilloway (2011) of 122 adolescents from 2 schools in the South of Ireland found that one quarter of victims did not confide in anybody. However, as only 25 respondents answered the question about whether they would report their experience or not to another individual, a broader sample of respondents is needed in order to have confidence that these results provide an accurate reflection of the general adolescent response pattern in relation to reporting cyberbullying.

The recently published HBSC report - 'Irish Health Behaviour in School-aged Children' (Kelly et al, 2012) – found that statistically significant differences exist by gender and age group - with more boys reporting having being bullied compared to girls and younger children more likely to report ever being bullied as compared to older children. These findings are particularly concerning in light of consistent evidence that girls tend to suffer more cyberbullying experiences than boys and that cyberbullying experiences tend to increase during adolescence. Whilst valuable in that it highlights age and gender distinctions regarding the self-reporting of bullying experiences in general, the HBSC measurement instrument does not provide the level of granularity necessary to determine the factors that are influencing adolescent resistance to report their cyberbullying experiences. Similarly, whilst providing evidence of adolescent resistance to report cyberbullying, O' Moore and Minton's (2011) study does not provide insight as to the causal reasons for that resistance. The authors speculate that the explanation for adolescent non-reporting may be a perception of greater self-efficacy than teacher efficacy in dealing with online problems or a lack of confidence in the school's abilities to deal with bullying (2011: 40). However, these are merely speculations, they are not empirically derived and simply are the interpretation of the authors. Consequently, neither study progresses our understanding of the factors underlying Irish adolescents' resistance to report cyberbullying experiences, nor provide insight as to whether individual or situational characteristics influence that resistance.

As previously noted, adolescent resistance to reporting is equally prevalent in the traditional (face-to-face) bullying context with evidence (Smith and Shu, 2000; Whitney and Smith, 1993) to show that 30-50% of pupils do not inform a parent or teacher that they had been a target of bullies. The influence of age on reporting behaviour is evident in Rigby and Slee's (1993) study which found that whilst approximately half of Australian students aged between 8 and 12 stated that they would like help

prevent others being bullied, as they increased in age, they became more reluctant to confide in or seek seeking adult intervention. However, the reasons underlying adolescents' reluctance to seek adult intervention or discuss the bullying experiences remain undetermined. Similarly, the degree to which gender, age or other variables apply in the case of Irish adolescents' resistance to report cyberbullying experiences has yet to be established.

5 CONCLUSIONS

The literature on cyberbullying is embryonic and as a consequence many deficits exist in relation to our understanding of the phenomenon. A growing body of evidence points to the fact that many adolescents who have been the target of cyberbullying behaviour choose to confide in peers rather than adult caregivers or teachers, despite having previously stated their intention to inform the latter should they themselves become the target of such behaviour. As a result, this behaviour remains unchecked and its impact unaddressed. The literature provides evidence that this chasm between reporting intention and actual behavioral outcome is consistent regardless of national differences. Despite this fact, remarkably little attention has been paid to understanding the factors underlying this resistance and the ensuring silence on the part of adolescents regarding their experiences of cyberbullying. Such insights are necessary if parents, teachers and those involved in the formulation of anti-bullying school policies are to be successful in their attempts to counter and eliminate cyberbullying behaviour. Research on the factors underlying adolescent resistance-to-report is therefore urgently mandated.

REFERENCES

Aftab, P. (2008). 'What methods work with the different kinds of cyberbullies?' [Online]. Available at: http://www.stopcyberbullying.org/educators/howdoyouhandleacyberbully.html. Accessed December 29th 2013.

Agatston, P. W., Kowalski, R. & Limber, S. (2007). Students' perspectives on cyber bullying. Journal of Adolescent Health, 41, 59-60.

Black, S., Weinles, D. and Washington E. (2010). Victim strategies to stop bullying. *Youth Violence and Juvenile Justice.* 8(2), 138-147.

Charach, A., Pepler, D., & Ziegler, S. (1995). Bullying at school: A Canadian perspective. *Education Canada* (Spring, 1995).

Cornell D. and Brockenborough, K. (2004). Identification of bullies and victims: A comparison of methods. Journal of School Violence 3(2-3).

Cotter, P. and McGilloway, S. (2011). Living in an 'Electronic Age': Cyberbullying among Irish Adolescents. *The Irish Journal of Education* 39 (44-56).

DeLara, E. (2008). Developing a philosophy about bullying and sexual harassment: Cognitive coping strategies among secondary school students. Journal of School Violence, 7, 72–96.

DeLara, E., (2012): Why Adolescents Don't Disclose Incidents of Bullying and Harassment, Journal of School Violence, 11:4, 288-305.

DES (Department of Education and Science), (n.d.). Anti-bullying policy.
Available at: http://www.education.ie/en/Publications/Education-Reports/Action-Plan-On-Bullying-2013.pdf.

Garbarino, J., & DeLara, E. (2002). And words can hurt forever: How to protect ado- lescents from bullying, harassment, and emotional violence. New York, NY: Simon & Schuster /The Free Press.

Goodstein, A. (2007). Totally wired: What teens and tweens are really doing online. New York: St. Martin's Griffin.

Holfeld, B. & Grabe, M. (2012). An examination of the history, prevalence, characteristics, and reporting of cyberbullying in the United States. In Q. Li, D. Cross & P.K. Smith (Eds.). Cyberbullying in the global playground: Research from international perspectives. Blackwell.

Irish Independent (2008). 'Cyberbullying "widespread" but parents kept in the dark' by Louise Hogan [Online]. Available at: http://www.independent.ie/national-news/cyber-bullying-widespread-but-parents-kept-in-the-dark-1380630.html.

Juvonen, J., & Gross, E. (2008). Extending the school grounds? Bullying experiences in cyberspace. *The Journal of School Health*, 78(9), pp. 496-505.

Kelly, C., Gavin, A., Molcho, M. & Nic Gabhainn, S. (2012). *The Irish Health Behaviours in School-aged Children (HBSC) study 2010.*Dublin: Department of Health.

Li, Q. (2008). A cross-cultural comparison of adolescents' experience related to.
cyberbullying. *Educational Research* Vol. 50, No. 3, September 2008, 223–234.

Li, Q. (2007). New bottle but old wine: A research on cyberbullying in schools. *Computers and Human Behaviour*, 23(4), 1777-1791.

Li, Q. (2006). Cyberbullying in Schools: A Research of Gender Differences. *School Psychology International, 27*(2), 157-170.

Naylor, P., & Cowie, H. (1999). The effectiveness of peer support systems in challenging school bullying: The perspectives and experiences of teachers and pupils. *Journal of Adolescence, 22*, 467–479.

NCH (2005). Putting U in the picture – mobile phone bullying survey 2005. Available at:

http://www.nch.org.uk/uploads/documents/Mobile_bullying_%20report.pdf.

Olweus, D. (1993). Bullying at School: What we know and what we can do. Oxford: Blackwell Publishing.

Olweus, D. (1999). Sweden. In Smith, P.K., Morita, Y., Junger-Tas, J., Olweus, D., Catalano, R., Slee, P. (Eds.), The Nature of School Bullying: A cross-cultural perspective. London: Routledge.

O'Higgins-Norman, J & Connolly, J. (2011). Mimetic Theory and Scapegoating in the Age of Cyberbullying: The Case of Phoebe Prince. Pastoral Care in Education: Vol.29: Issue 4. London: Routledge.

O'Moore, A. M. & Minton, S. J. (2011). Cyber-Bullying: The Irish Experience. New York: Nova Science Publishers, Inc.

Palfrey, J. & Gasser, U. (2008). *Born Digital: Understanding the First Generation of Digital Natives.* New York: Basic Books.

Pellegrini, A. and Bartini, M. (2000). A Longitudinal Study of Bullying, Victimizatioin, and Peer Affiliation During The Transistion From Primary School To Middle School, *American Educational Research Journal* 37(3): 699–725.

Pepler, D., Jiang, D., Craig, W., & Connolly, J. (2008). Developmental trajectories of bullying and associated factors. Child Development, 79, 325–338.

Petrosino, A., Guckenburg, S., DeVoe, J., & Hanson, T. (2010). What characteristics of bullying, bullying victims, and schools are associated with increased reporting of bullying to school officials? (Issues & Answers Report, REL 2010, No. 092). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Rigby, K. and Slee, P.T. (2008). Interventions to reduce bullying, International Journal of Adolescent Medicine and Health, 20,165-183.

Shariff, S. (2008). Cyberbullying: Issues and solutions for the school, the classroom and the home. New York: Routledge.

Slonje, R. and Smith, P.K. (2008). Cyberbullying: Another main type of bullying? Scandinavian Journal of Psychology, 49(2), pp. 147-154.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S.,Russell, S., Tippett N. (2008). Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry 49:4 (2008), pp 376–385.

Smith, P., et al. (2006). UK Report: *An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying.* Research Brief No. RBX03-06. London: DfES.

Smith, P. K. and Shu, S. (2000). 'What Good Schools Can Do About Bullying: findings from a survey in English schools after a decade of research and action', in 'Childhood', Vol 7 (2).

Tannen, D. (2004). He Said, She Said: Gender, Language and Communication. [Video]. Narrated by Deborah Tannen. California: Modern Scholar.

Tannen, D. (1995). Talking from Nine to Five: Women and Men in the Workplace: Language, Sex and Power. New York: Harper Collins.

Tannen, D. (1991). You Just Don't Understand: Women and Men in Conversation. New York: Ballantine Books.

Trach, J., et al. (2012). To Bully or Not to Bully, That Is Not the Question: Western Australian Early Adolescents' in Search of a Reputation. Journal of Adolescent Research July 1, 2012 27: 498-522.

Trach, J. Hymel, S., Waterhouse, T., & Neale, K. (2010). Age Differences in Bystander Responses to School Bullying: A Cross-Sectional Investigation. Canadian Journal of School Psychology, 25 (1), 114-130.

Trolley B., Hanel C. and Shields L. (2006). Demystifying & Deescalating Cyber Bullying in the Schools: A Resource Guide for Counselors, Educators & Parents. Book Locker, Bangor, ME.

Wall Street Journal (2007). Article: *In Study, Facebook Users Share with Stranger.* August 14th. http://online.wsj.com/article/SB11870598778049.

Whitney, I. & Smith, P. K. (1993). A survey of the nature and extent of bullying in junior/middle and secondary schools. *Education Research*, 35, 3-25.

Willard, N. (2007). Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Illinois: Research Press.

Ybarra, M. L., Mitchell, K. J., Wolak, J., Finkelhor, D. (2006). Examining Characteristics and Associated Distress Related to Internet Harassment: Findings From the Second Youth Internet Safety Survey. *Pediatrics* 2006; 118;1169-1177.

Ybarra, M., & Mitchell, K. (2004). Youth engaging in online harassment: associations with caregiver-child relationships, Internet use, and personal characteristics. *Journal of Adolescence, 27*(3), 319-336.

Electronic Health Record Portal Adoption by Health Care Consumers
Proposal of a New Adoption Model

Jorge Tavares and Tiago Oliveira
ISEGI, Universidade Nova de Lisboa, Rua Campolide, Lisboa, Portugal
tavares.jorge@gmail.com, toliveira@isegi.unl.pt

Keywords: e-Health, Technology Adoption, UTAUT2, Health Care Consumers, e-Government.

Abstract: The aim of this study is to identify a set of determinants of adoption of electronic health records (EHR) portals by health care consumers. Based on extensive literature review we suggest a new research model based on the unified theory of acceptance and use of technology in a consumer context (UTAUT2) by integrating a second order construct, Concern for Information Privacy (CFIP) framework and a moderator, chronic disability. A set of propositions is also included to test the new conceptual model. We also present a plan to validate the proposed model through empirical testing. The EHR portals are a part of the e-government strategy currently unfolding in Portugal. Understanding the acceptance and use of EHR portals by health care consumers should benefit the future sustainability of the Heath Care System, which will gain a more efficient use of resources.

1 INTRODUCTION

E-health technology for health care consumers is the use of electronic resources, mainly web-based, on medical topics by healthy individuals or patients (Alpay et al., 2010; Lee et al., 2010; Millard and Fintak 2002; Nazi, 2003). Our study focuses on a specific type of e-health technology, the electronic health records (EHR) portals, which give patients access to medical records, exam results, and services, such as appointment scheduling, notification systems, and e-mail access to the doctor (Angst and Agarwal, 2009, Andreassen et al., 2007).

In Portugal, where data collection for this study will be implemented, approximately 30% of the population already uses the internet for health purposes (Andreassen et al., 2007), considerabl less than Northern European countries like Denmark and Norway, where more than 50% of the population uses Internet for health purposes (Andreassen et al., 2007). Understanding the acceptance and use of EHR portals technology by health care consumers is a very important topic with clear benefits for society and the future sustainability of the Heath Care System (Or and Karsh, 2009; Wilson and Lankton, 2004). In a country like Portugal, which faces a severe austerity program, the cost and efficiency advantages of e-health technology use by patients brings benefits to health care consumers and relief to the Ministry of Health budget (McKee et al., 2012; Metaxiotis et al., 2004). The EHR portals are a specific type of technology that can greatly help to achieve these benefits for both patients and health care providers (Angst and Agarwal, 2009). EHR portals are an initiative promoted by the Portuguese government that is a part of a broader e-government strategy seeking to facilitate services and communications between public services and the citizens. The most important initiative is the "Portal do Utente" (User's Portal) a national EHR portal, created by the Ministry of Health that allows all Portuguese citizens to schedule appointments with their general practitioner, obtain electronic medical prescriptions, access medical records and exam results, and share information with health care providers (Rodrigues et al., 2013; Pereira, 2012). With more than half a million users already registered, the objective of the Ministry of Health is to make this portal a primary point of contact between the patient and the national health care system that in Portugal provides coverage to the entire population (Rodrigues et al., 2013; Pereira, 2012).

Understanding whether or not the current adoption models are suitable to a particular field of

interest, by verifying and proposing the necessary changes and extensions (Alvesson and Kaerreman, 2007; Martins et al., 2014; Miltgen et al., 2013) is a vital topic in the area of information system (IS) acceptance and use (Venkatesh et al., 2012). In our study we will examine these assumptions in the field of e-health technology use and acceptance by the health care consumers and propose a new research model based on the unified theory of acceptance and use of technology in a consumer context (UTAUT2) by integrating a second-order construct based on the Concern for Information Privacy (CFIP) framework and a new moderator, chronic disability (Angst and Agarwal 2009; Millard and Fintak, 2002).

The paper proceeds as follows: First we review the literature concerning information technology (IT) adoption models regarding consumer health care. We then present a research model to analyze ERHs portals for the health care consumer. Finally we discuss the issue and present conclusions.

2 THEORETICAL BACKGROUND

There have been several theoretical models developed from theories in psychology, sociology, and consumer behavior employed to explain technology acceptance and use (Venkatesh et al., 2012). The goal of this study is to focus specifically on the e-health adoption in the perspective of the health care consumer, so it is of the utmost importance to review the literature in this particular field. Adoption of e-health technologies by the patients is clearly a very important topic in IS. The adoption of e-health technologies by health care professionals still requires more attention and research due to the limited number of studies reported in the literature to date (Angst and Agarwal, 2009; Or and Karsh, 2009).

When studying e-health and health care adoption by health care professionals the most common adoption models used are the technology acceptance model (TAM) (Dunnebeil et al., 2012; Ketikidis et al., 2012) and UTAUT (Chang et al., 2007; Yi et al., 2006). Evaluating the studies published in the field of consumer health information technology adoption and more specifically in the use and adoption of e-health tools by the health care consumer, most studies use TAM or extensions of TAM (Or and Karsh, 2009; Wilson and Lankton, 2004). TAM was designed and tailored in IS contexts to predict information technology acceptance and usage on the job (Venkatesh et al., 2003). Independent attempts by several researchers to expand TAM in order to adapt it to the constantly changing IT environments, like e-health, has led to a state of theoretical chaos and confusion in which it is not clear which version of the many iterations of TAM is the commonly accepted one (Wilson and Lankton, 2004; Benbasat and Barki, 2007). UTAUT formulate a unified model that integrates elements of eight models in the field of IT acceptance. The eight models are the theory of reasoned action (Fishbein and Ajzen, 1975); the technology acceptance model (Davis, 1989); the motivational model (Davis et al., 1992); the theory of planned behavior (Ajzen, 1991); a model combining the technology acceptance model and the theory of planned behavior (Taylor and Todd, 1995); the model of PC utilization (Thompson et al., 1991); the innovation diffusion theory (Rogers, 1995), and the social cognitive theory (Compeau et al., 1999; Compeau and Higgins, 1995). The R^2 obtained with UTAUT was superior to any of the individual models including TAM (Venkatesh et al., 2003).

Although UTAUT provides better results than TAM and other IS adoption models, the focus of UTAUT is also the employee technology acceptance at individual level (Venkatesh et al., 2003; Venkatesh et al., 2012). Ideally, we then need a model tailored to the consumer use context, and in this specific field UTAUT2 was developed with this goal, obtaining excellent results (Venkatesh et al., 2012). When compared with UTAUT in the consumer use context, the new UTAUT2 model obtained higher R^2, and is better able to explain the reasons behind the adoption (Venkatesh et al., 2012). If UTAUT2 outperformed UTAUT in the consumer use context, we believe that UTAUT2 could be used as the standard starting model for studying e-health adoption.

E-health technology adoption differentiates from IT adoption in general due to the sensitive topics and issues related to health status of an individual, making the drivers of adoption in e-health different from other IT technologies (Angst and Agarwal, 2009). In health care, confidentiality is the ethical principle that a health professional will keep confidential all information relating to a patient, unless the patient consents to disclosure (Bauer, 2002). Several studies point out that awareness of lack of confidentiality and privacy concerns may reduce the adoption of e-health tools by the patients and health care consumers (Angst and Agarwal, 2009; Fisher and Clayton, 2012; Fogel and Nehmad, 2009; O'Donnell et al., 2011). Studies focusing specifically on EHR portals show that patients are

highly concerned about privacy of their personal medical records (Angst and Agarwal, 2009). Literature review identified that patients with chronic illness, severe illness, or disability are more likely to use e-health technologies if they have the resources and support available. It is paramount to understand that it is only if patients who are chronically or severely disabled have the conditions and resources available, will they access the health record portals. (Millard and Fintak, 2002; Renahy et al., 2008).

3 RESEARCH MODEL

UTAUT2 was developed as an adoption model providing the general factors of IT adoption in the consumer use. However, according to Venkatesh et al. (2012) in certain situations where the technology may be influenced by specific factors it may be necessary to make extend the model with new constructs, moderators, and relationships. We therefore identified key additional constructs and relationships based on the literature review that are specific to IT health care adoption to be integrated into UTAUT2, thus tailoring it to e-health consumer context, more specifically to study the adoption of EHR portals. We did this by (1) identifying key constructs from earlier research in IT health care consumer adoption (confidentiality) and by (2) adding a new moderator specific to IT health care use (chronic disability). Figure 1 illustrates the new research model.

3.1 UTAUT2 Model

Performance expectancy is defined as the degree to which using a technology will provide benefits to consumers in carrying out certain activities (Venkatesh et al., 2003). Literature review indicates that health care consumers tend more to adopt e-health technologies that provide clear benefits, such as obtaining an electronic medical prescription via EHR portals (Alpay et al., 2010; Arsand and Demiris, 2008; Keselman et al., 2008).

P1. Performance Expectancy will Positively Influence Behavioral Intention.

Effort expectancy is the degree of ease related to consumers' use of technology (Venkatesh et al., 2003). The easier that consumers can understand and use an e-health technology, the greater is the probability that they will adopt it (Alpay et al., 2010; Keselman et al., 2008).

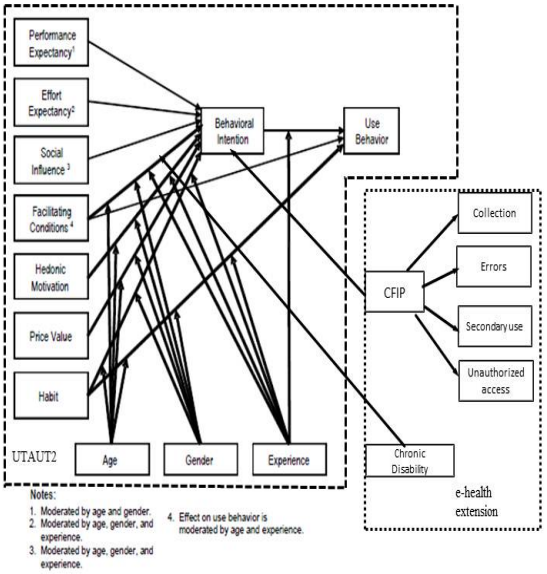


Figure 1: The Research Model.

P2. Effort Expectancy will Positively Influence Behavioral Intention.

Social influence is the extent to which consumers perceive that others who are important to them (e.g., friends and family), believe they should use a particular technology (Venkatesh et al., 2012). In the case of e-health this can be also an important construct, since people that share the same diseases (e.g., Multiple Sclerosis) or the same health condition (e.g., obesity) tend to be influenced by what others in the same conditionuse and do (Fisher and Clayton, 2012; Thackeray et al., 2013).

P3. Social Influence will Positively Influence Behavioral Intention.

Facilitating conditions refer to consumers' perceptions of the resources and support available to execute a behavior (Venkatesh et al., 2003). In the specific case of e-health this construct can be very important due to the fact that patients and disabled people may have special needs that healthy people do not (Arsand and Demiris, 2008; Millard and Fintak, 2002).

P4. E-health Platforms with Greater Resources and Support Available (Facilitating Conditions) are more likely to Positively Influence Behavioral Intention and Use Behavior.

Hedonic motivation is defined as intrinsic motivation (e.g., enjoyment) and has been included as a key predictor in much of the reported consumer behavior research (Venkatesh et al., 2012). Dealing and obtaining information about our health status by using e-health technologies may be a rewarding

process, or in some cases may not be when a patient is dealing, for example, with cancer (Lee et al., 2010).

P5. Hedonic Motivation will Positively or Neutrally Influence Behavioral Intention.

Price value: in a consumer use environment, price is also a relevant factor, as unlike workplace technologies, consumers must bear the costs related with the purchase of devices and services (Venkatesh et al., 2012). If a patient can obtain his medical prescription via an EHR portal, he can save transportation money by avoiding a trip to a health center or hospital. The better the perception a health care consumer has about the "price value" of an e-health technology (i.e., that it can help him save money), the more likely it is that he will adopt it (Alpay et al., 2010; Metaxiotis et al., 2004).

P6. Price Value will Positively Influence Behavioral Intention.

Habit can be defined as the extent to which people tend to execute behaviors automatically because of learning (Venkatesh et al., 2012). We can expect that habit will positively influence e-health adoption, as it does in other IT adoption fields, since habit is a concept that should not be specific to an IT technology (Venkatesh et al., 2012).

P7. Habit will Influence Positively Behavioral Intention and use Behavior.

Age, gender and experience are theorized to moderate various UTAUT2 relationships (Venkatesh et al., 2012). Literature review indicates that in IT technologies in general and specifically in e-health, younger people and women tend to have the habit to use more e-health technologies (Millard and Fintak, 2002; Thackeray et al., 2013; Venkatesh et al., 2012). According to the literature review, people with less experience in using a technology tend to be more influenced by experience (Venkatesh et al., 2012).

P8. Age and Gender will Moderate the Effect of Habit on Behavioral Intention, such that the Effect will be Stronger for Younger Women.

P9. Age and Gender will Moderate the Effect of Habit on Use Behavior, such that the Effect will be Stronger for Younger Women.

P10. Experience will Moderate the Effect of Behavioral Intention on Use, such that the Effect will be Stronger for Consumers with Less Experience.

3.2 e-Health Extension

Smith, Milberg, and Burke (1996) developed and tested the CFIP construct to measure attitudes and beliefs about individual information privacy related to the use of personal information in a business setting. They conceptualized CFIP as being composed of four distinct, yet correlated, latent factors: collection, errors, unauthorized or improper access, and secondary use. Collection is the concern that an extensive amount of personal information is being collected and stored in databases (Smith et al., 1996). Errors are directly linked with the concern that protection against deliberate and accidental error in personal data is inadequate (Smith et al., 1996). Unauthorized access is the concern that data about individuals is available to people not authorized to view or work with these data (Smith et al., 1996). Secondary use is the apprehension that information is collected from individuals for one purpose but is used for another secondary purpose without authorization from the individuals (Smith et al., 1996). Stewart and Segars (2002) expanded upon the Smith et al. (1996) study and not only validated the multidimensional nature of the CFIP construct, but also found support for the hypothesis that a second-order factor structure is empirically valid, thus settling the complexity of an individual's concern for information privacy with direct influence in the behavioral intention to use a technology (Stewart and Segars, 2002). Angst et al. (2009) studied the adoption of EHRs in the presence of privacy and used in their study the CFIP framework as a second order factor structure. To measure CFIP Angst et al. (2009) adapted the scale developed by Smith et al. (1996). Minor changes were made to their instrument to reflect privacy concerns relative to health data instead of corporate data. Based on the earlier use of the CFIP framework in e-health, we add CFIP as a predictor of health consumer behavioral intention to use a technology.

P11. CFIP will Positively Influence Behavioral Intention.

Chronic disability is an incapacitating situation (e.g., chronic illness) that affects a patient permanently or for long-term periods. Review of literature points out that patients with chronic illness or disability are more likely to use e-health technologies if they have the resources and support available (facilitating conditions) (Millard and Fintak, 2002; Thackeray et al., 2013).

P12. We can Theorize that Chronic Disability is a

Positive Moderator of Facilitating Conditions to Explain Behavioral Intention.

4 METHODOLOGICAL IMPLICATIONS AND VALIDATION APPROACHES

To test the propositions suggested here, we will develop a pilot study with a sample of 30 users of EHR portals. The final questionnaire will be administered in Portuguese. To ensure translation equivalence we will translate the English questionnaire to Portuguese and then back to English; these tasks will be executed by professional translators (Brislin, 1970). We will test the hypothesized relationships among the constructs using structural equation modeling (SEM). There are two families of SEM techniques (Henseler et al., 2009): covariance-based techniques and variance-based techniques. Partial least squares (PLS) is a variance-based technique and will be used in this research since, (i) the research model has not been tested in the past; (ii) the research model is considered complex. The theoretical approach developed in this study should be implemented first in Portugal focusing in the EHR portals. The data will be collected in EHR portals managed by the Portuguese Ministry of Health, allowing a vast coverage of the Portuguese population since Portugal has a National Health System that covers all of the population. The eligibilities criteria's to answer the questionnaire are: Portuguese nationality and age equal to or greater than 18 years old. Our objective is to collect a final sample with more than 300 questionnaires.

5 LIMITATIONS AND FUTURE RESEARCH

This study is in its first theoretical concept, in which an initial model is suggested based on the literature review and conceptual reasoning. The next step is the application and validation of the model to a set of health care consumers, in order to test the developed framework and directly assess its explanatory and predictive power. Future studies may evaluate other relationships that were not foreseen in this model and that will improve the ability to explain the dependent variables. Thus, this study opens up other options for future research. Refinement of the constructs and measures is one of the possibilities. Another option is the investigation of more complex relationships between the independent and dependent variables of the model. Testing this model with other e-health technologies and in other countries that may be more or less developed than Portugal in e-health use are options that can also bring added value.

6 CONCLUSIONS

Understanding the acceptance and use of EHR portals by health care consumers should bring strong benefits for the future sustainability of the Heath Care System, which will enjoy more efficient use of resources. Thus, the aim of this study is to identify a set of determinants of adoption of EHR portals by health care consumers. To realize this goal we suggest a research model based on UTAUT2, adding new constructs (CFIP) and a new moderator of chronic disability. We designate this new set of constructs and the new moderator "e-health extension to UTAUT2". We also expect this study to provide a theoretical framework that is a foundation and a starting point for future research on the adoption of e-health by health care consumers.

REFERENCES

Ajzen, I. (1991) 'The Theory of Planned Behaviour', *Organizational Behavior and Human Decision Processes,* 50(2), 179-211.

Alpay, L. L., Henkemans, O. B., Otten, W., Rovekamp, T. A. J. M. and Dumay, A. C. M. (2010) 'E-health Applications and Services for Patient Empowerment: Directions for Best Practices in The Netherlands', *Telemedicine Journal and E-Health,* 16(7), 787-791.

Alvesson, M. and Kaerreman, D. (2007) 'Constructing mystery: Empirical matters in theory development', *Academy of Management Review,* 32(4), 1265-1281.

Andreassen, H., Bujnowska-Fedak, M., Chronaki, C., Dumitru, R., Pudule, I., Santana, S., Voss, H. and Wynn, R. (2007) 'European citizens' use of E-health services: A study of seven countries', *BMC Public Health,* 7(1), 53.

Angst, C. M. and Agarwal, R. (2009) 'Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and Individual Persuasion', *MIS Quarterly,* 33(2), 339-370.

Arsand, E. and Demiris, G. (2008) 'User-centered methods for designing patient-centric self-help tools', *Informatics for Health & Social Care,* 33(3), 158-169.

Bauer, K. A. (2002) 'Using the Internet to empower patients and to develop partnerships with clinicians',

World hospitals and health services : the official journal of the International Hospital Federation, 38(2), 2-10.

Benbasat, I. and Barki, H. (2007) 'Quo vadis, TAM?', *Journal of the Association for Information Systems,* 8(4), 212-218.

Brislin, R. W. (1970) 'Back-Translation for Cross-Cultural Research', *Journal of Cross-Cultural Psychology,* 1(3), 185-216.

Chang, I. C., Hwang, H.-G., Hung, W.-F. and Li, Y.-C. (2007) 'Physicians' acceptance of pharmacokinetics-based clinical decision support systems', *Expert Systems with Applications,* 33(2), 296-303.

Compeau, D., Higgins, C. A. and Huff, S. (1999) 'Social cognitive theory and individual reactions to computing technology: A longitudinal study', *MIS Quarterly,* 23(2), 145-158.

Compeau, D. R. and Higgins, C. A. (1995) 'Application of Social Cognitive Theory to Training for Computer Skills', *Information Systems Research,* 6(2), 118-143.

Davis, F. D. (1989) 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology', *MIS Quarterly,* 13(3), 319-340.

Davis, F. D., Bagozzi, R. P. and Warshaw, P. R. (1992) 'Extrinsic and Intrinsic Motivation to Use Computers in the Workplace', *Journal of Applied Social Psychology,* 22(14), 1111-1132.

Dunnebeil, S., Sunyaev, A., Blohm, I., Leimeister, J. M. and Krcmar, H. (2012) 'Determinants of physicians' technology acceptance for e-health in ambulatory care', *International Journal of Medical Informatics,* 81(11), 746-760.

Fishbein, M. and Ajzen, I. (1975) *Belief, attitude, intention, and behavior : an introduction to theory and research,* Reading, Mass.: Addison-Wesley.

Fisher, J. and Clayton, M. (2012) 'Who Gives a Tweet: Assessing Patients' Interest in the Use of Social Media for Health Care', *Worldviews on Evidence-Based Nursing,* 9(2), 100-108.

Fogel, J. and Nehmad, E. (2009) 'Internet social network communities: Risk taking, trust, and privacy concerns', *Computers in Human Behavior,* 25(1).

Henseler, J., Ringle, C. M. and Sinkovics, R. R. (2009) 'The use of partial least squares path modeling in international marketing.' in (Eds)., I. R. R. S. a. P. N. G., ed. New Challenges to International Marketing, Stamford: Jai Press Inc., 277-319.

Keselman, A., Logan, R., Smith, C. A., Leroy, G. and Zeng-Treitler, Q. (2008) 'Developing Informatics Tools and Strategies for Consumer-centered Health Communication', *Journal of the American Medical Informatics Association,* 15(4), 473-483.

Ketikidis, P., Dimitrovski, T., Lazuras, L. and Bath, P. A. (2012) 'Acceptance of health information technology in health professionals: An application of the revised technology acceptance model', *Health Informatics Journal,* 18(2), 124-134.

Lee, C.-j., Gray, S. W. and Lewis, N. (2010) 'Internet use leads cancer patients to be active health care consumers', *Patient Education and Counseling,* 81S, S63-S69.

Martins, C., Oliveira, T. and Popovič, A. (2014) 'Understanding the Internet banking adoption: An unified theory of acceptance and use of technology and perceived risk application.', *International Journal of Information Management*, 34(1), 1-13.

McKee, M., Karanikolos, M., Belcher, P. and Stuckler, D. (2012) 'Austerity: a failed experiment on the people of Europe', *Clinical Medicine,* 12(4), 346-350.

Metaxiotis, K., Ptochos, D. and Psarras, J. (2004) 'E-health in the new millennium: a research and practice agenda', *International journal of electronic healthcare,* 1(2), 165-75.

Millard, R. W. and Fintak, P. A. (2002) 'Use of the Internet by patients with chronic illness', *Disease Management & Health Outcomes,* 10(3).

Miltgen, C. L., Popovič, A. and Oliveira, T. (2013) 'Determinants of end-user acceptance of Biometrics: Integrating the "Big 3" of technology acceptance with privacy context.', *Decision Support Systems,* 53, 103-114.

Nazi, K. M. (2003) 'The journey to e-Health: VA Healthcare Network Upstate New York (VISN 2)', *Journal of medical systems,* 27(1), 35-45.

O'Donnell, H. C., Patel, V., Kern, L. M., Barron, Y., Teixeira, P., Dhopeshwarkar, R. and Kaushal, R. (2011) 'Healthcare Consumers' Attitudes Towards Physician and Personal Use of Health Information Exchange', *Journal of General Internal Medicine,* 26(9).

Or, C. K. L. and Karsh, B.-T. (2009) 'A Systematic Review of Patient Acceptance of Consumer Health Information Technology', *Journal of the American Medical Informatics Association,* 16(4), 550-560.

Pereira, G. (2012) 'Meio milhão de utentes já marca consultas pela Internet - JN', *Jornal de Notícias*, available: http://www.jn.pt/PaginaInicial/Sociedade/Saude/Interior.aspx?content_id=2723506 [accessed 18 Oct 2012].

Renahy, E., Parizot, I. and Chauvin, P. (2008) 'Health information seeking on the Internet: a double divide? Results from a representative survey in the Paris metropolitan area, France, 2005-2006', *BMC Public Health,* 8(1), 69.

Rodrigues, D. F., Lopes, J. C. and Tavares, J. F. (2013) *Maninfold Marketing: A New Marketing Archetype for the Information Age, Applied to the Adoption of Oral Contraceptives and Other Drugs by End-Users., Proceedings of the Third Annual Conference of International Network of Business & Management Journals* (INBAM 2013), Lisbon.

Rogers, E. (1995) *Diffusion of Innovations,* Free Press, New York.

Smith, H. J., Milburg, S. J. and Burke, S. J. (1996) 'Information privacy: Measuring individuals' concerns about organizational practices', *MIS Quarterly,* 20(2), 167-196.

Stewart, K. A. and Segars, A. H. (2002) 'An empirical examination of the concern for information privacy

instrument', *Information Systems Research,* 13(1), 36-49.

Taylor, S. and Todd, P. (1995) 'Assessing IT usage: The role of prior experience', *MIS Quarterly,* 19(4), 561-570.

Thackeray, R., Crookston, B. T. and West, J. H. (2013) 'Correlates of health-related social media use among adults', *Journal of medical Internet research,* 15(1), e21.

Thompson, R. L., Higgins, C. A. and Howell, J. M. (1991) 'Personal Computing: Toward a Conceptual Model of Utilization', *MIS Quarterly,* 15(1), 125-143.

Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. (2003) 'User acceptance of information technology: Toward a unified view', *MIS Quarterly,* 27(3), 157-178.

Venkatesh, V., Thong, J. Y. L. and Xu, X. (2012) 'Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology', *MIS Quarterly,* 36(1), 425-478.

Wilson, E. V. and Lankton, N. K. (2004) 'Modeling patients' acceptance of provider-delivered e-health', *Journal of the American Medical Informatics Association,* 11(4), 241-248.

Yi, M. Y., Jackson, J. D., Park, J. S. and Probst, J. C. (2006) 'Understanding information technology acceptance by individual professionals: Toward an integrative view', *Information & Management,* 43(3), 350-363.

Internet Science for Strategic Planning

Žiga Turk
University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova 2, Ljubljana, Slovenia
ziga.turk@fgg.uni-lj.si

Keywords: Information Technology, Internet Science, Strategic Planning, Future Studies, Policy Creation.

Abstract: The rapid development of information and communication technology (ICT) is perhaps the most influential driver that is fundamentally changing the world and the societies we live in. ICT is (a) changing the communication fabric that is linking the elements of societies together and is (b) automating human routine work. The latter is enabling automation and creating an abundance of food, industrial products and information. This abundance is pushing the value creation towards the creation of new knowledge and meaningful (rather than only functional) products and services. Information, knowledge and meaning are the three key commodities of the modern economy. Innovation and creativity are key processes creating these commodities. The two activities are very significantly supported by information and communication technologies.

ICTs are politically acknowledged on several levels of future planning: in R&D programs, development strategies, future studies and visions. However, a scientific base for all this is lacking. The interaction among ICT, innovation and society at large is a subject of research projects such as EINS and spawning the birth of a new interdisciplinary science, Internet Science - that is studying the interaction between information technology and society.

1 INTRODUCTION

The history of humanity shows a very clear trend that an increasing number of people is involved in the information professions - that is, the result of their work is an information; not an agricultural or an industrial product. They are sometimes referred as "knowledge workers".

The growing number of the knowledge workers vis-a-vis industrial and agricultural workers is a quantifiable indicator of society's transition from agricultural or industrial society into what is called "the information society". The growing contribution of the information/knowledge/meaning to the value of a product is another indicator of the transition of the society from the material towards the informational.

There are some who are charting the trajectory forward; from information society towards knowledge society and society of meaning (Pink, 2006) or emphatic society (Rifkin, 2009). However, from the perspective of analysing processes and technologies these are very similar to information society.

The key commodity in the information society is information. Information is created, communicated and consumed. All these processes are supported by media, tools and technologies. For example, an old technology for communicating information is sending around paper. A modern way is using what we call ICTs by which we usually mean the electronic and digital tools and services.

Communication technology has, too date, undergone two major and two minor revolutions (Turk, 2009):

- The first minor revolution was the invention of paper, clay and other writing media thousands of years BC.
- The first major revolution was the democratization of this invention by the introduction of cheap paper and printing press in Europe some 500 years ago.
- The second minor revolution was the invention of electronic communication like telegraph, telephone, radio and TV in the late 19th and early 20th century.
- The second major communication revolution was the democratization of electronic communication of texts, sounds and videos using the internet.

While the impact of the minor revolutions was significant, it was limited. The impact of the major revolutions brought paradigm shifts to society as a

whole.

The ICTs are a key enabling technologies of the so-called knowledge triangle, which includes research, education and innovation, the three key activities of the "information (or knowledge or meaning) society".

Better linking of the three apexes of the triangle has been one of the goals of original as well as of the Lisbon Strategy (European Commission, 2009). The goal of that strategy was "to make Europe the most competitive and dynamic knowledge based economy in the world". It was based on the Schumpeterian assumption that innovation is the key driver of economic growth, and placed a big emphasis on innovation, research and the knowledge economy. The Bologna reform of the European Universities that was taking place at the roughly same time addressed the education apex of the triangle. A strong impetus towards the innovation union was also the Aho Report (Aho, 2006).

2 ICT IN POLICY PLANNING

2.1 Programs and Strategies

The ICT is broadly understood as a political development priority, in the EU at least since the Bangeman Report (European Union, 1994). ICT was prominently represented in the Lisbon Strategy. In the Europe 2020 strategy, Digital Agenda for Europe (European Commission, 2010) is one of the seven flagship projects. "The overall aim of the Digital Agenda is to deliver sustainable economic and social benefits from a digital single market based on fast and ultra-fast internet and interoperable applications."

In Project Europe 2030 (Reflection Group, 2010) the economic importance of ICT is well understood: "The EU must strengthen the Single Market against temptations of economic nationalism and complete it to include services, the digital society and other sectors, which are likely to become the main drivers of growth and job creation in a market of 500 million users and consumers" … "Digitalisation is increasing the scope for outsourcing, and the ICT-revolution may give ample scope for growth in productivity for decades to come." Also, the broader impacts of the ICT are mentioned: "The availability of multiple entry points through which citizens can engage with the legislative process should provide the EU with the legitimacy and flexibility it needs to address the challenges arising from globalization in a digital, interdependent, network-oriented and open society" … "… more systematic use of digital resources for e-governance". Role of ICT in research was also acknowledged by the European Council (2009).

2.2 ICT in Foresight Studies

ICT is one of the most vibrant sectors of the economy. It is therefore a subject of interest of the financial sector. Also, since the sector is advancing so rapidly it is opening up new issues with respect to regulation.

The European Internet Foundation made some predictions of the Industry by 2025 (European Internet Foundation, 2009). For 2025 they predict the "world driven by mass collaboration". But warn: "By 2025, the "digital arms race" between those intent on harm and those collaborating to prevent it will have become a central feature of our justice and law-enforcement systems".

A high level group prepared Future Internet 2020 (European Commission, 2009a) vision: "By 2020 the Internet will be both laid out as public infrastructures and dynamically created by the objects connecting to one another. We need to see the Internet of the Future as this seamless fabric of classic networks and networked objects. The content and services they facilitate will be all around us, always on, everywhere, all the time. It will lead the way to opportunities we never knew existed: new ways of working; new ways of interacting; new ways to be entertained; new ways of living. Next to these future applications today's Internet will look clunky and primitive. For instance, multimedia applications will move towards the bandwidth of human perception and beyond."

The bandwidth of human perception is estimated at 500 Mbits/seconds which is easily met today by the fiber to the home technology. Due to the limitations of the human body there will be no need for faster individual connections.

Several projects on the future of the internet take place in the EU, the US and beyond. Towards the Future Internet Project (Forge et.al., 2009) identified four possible future scenarios: "1. Smooth Trip - the rise of the internet economy as a whole life and work style - a middle road in contrast to more disruptive scenarios. 2. Going Green – internet technologies are used to combat growing environmental challenges. 3. Commercial Big Brother – a heavily commercialized consumer platform. 4. Power to the People – a forum for democracy and freedom."

The Paradiso Project (2011) had an objective "to explore how might or should our societies evolve in the next decades (a probable paradigm shift) and to derive from this analysis how can Information and Communication Technologies (ICT), and the Future Internet in particular, contribute to making this future better."

Internet technologies are perhaps most speculated about. Gartner invented and keep publishing the so called Gartner Hype Cycle that is showing technology maturity levels. They are also making shorter and medium term predictions (Gartner, 2010) of the industry.

There is no mention of the ICT in the "The World in 2025 Report" (European Commission, 2009b). The "Transformed World" (National Intelligence Council, 2008) report mentions possible technological leapfrogging of the developing world, internet as means of spreading propaganda and organizing NGOs, ubiquitous computing, role of ICT in warfare.

2.3 Visions of ICT

The various visions of ICT development are largely influenced by the fact that the industry keeps progressing according to Moore's law. Roughly doubling capacity and speeds every 18th months. This means an order of magnitude, tenfold, increase in less than 5 years. No other industry follows this cure, but an increasing number of industries are hooked on this curve by getting digitized: photography, music, movies, books, media … anything that is information.

Zittrain (2008) paints a worried picture that the future of the Internet may be as a number of closed networks dominated by commercial players like Apple and Facebook instead of an open collaboration platform. He writes: "In the arc from the Apple II to the iPhone, we learn something important about where the Internet has been, and something more important about where it is going. The PC revolution was launched with PCs that invited innovation by others. So too with the Internet. Both were generative: they were designed to accept any contribution that followed a basic set of rules (either coded for a particular operating system, or respecting the protocols of the Internet). Both overwhelmed their respective proprietary, non-generative competitors, such as the makers of stand-alone word processors and proprietary online services like CompuServe and AOL. But the future unfolding right now is very different from this past. The future is not one of generative PCs attached to a generative network. It is instead one of sterile appliances tethered to a network of control." but concludes "Our fortuitous starting point is a generative device in tens of millions of hands on a neutral Net. To maintain it, the users of those devices must experience the Net as something with which they identify and belong. We must use the generativity of the Net to engage a constituency that will protect and nurture it. That constituency may be drawn from the ranks of a new generation able to see that technology is not simply a video game designed by someone else, and that content is not simply what is provided through a TiVo or iPhone."

Among the works on the impact of the internet on politics and democracy, Morozov (2010) stands out with a cyber-realistic view that Internet does not necessarily bring democracy and freedom as some cyber-utopians would claim.

Kurzweil (1998) predicted in 1998 for 2009: "Personal computers with high-resolution visual displays come in a range of sizes, from those small enough to be embedded in clothing and jewellery up to the size of a thin book. Cables are disappearing. Communication between components uses short-distance wireless technology. High-speed wireless communication provides access to the Web."

Kurzweil (ibid.) for 2019: »A $1,000 computing device (in 1999 dollars) is now approximately equal to the computational ability of the human brain; Automated driving systems are now installed in most roads. There are widespread reports of computers passing the Turing Test, although these tests do not meet the criteria established by knowledgeable observers.«

Kurzweil (ibid.) for 2029: »A $1,000 (in 1999 dollars) unit of computation has the computing capacity of approximately 1,000 human brains; There is almost no human employment in production, agriculture, or transportation. Basic life needs are available for the vast majority of the human race. There is a growing discussion about the legal rights of computers and what constitutes being "human. Machines claim to be conscious. These claims are largely accepted."

And, finally, for 2099: »There is no longer any clear distinction between humans and computers; Life expectancy is no longer a viable term in relation to intelligent beings.«

3 ICT AND SOCIETY

The interactions are too numerous to analyse in this

context as the ICT penetrates all aspects of society. One report that stands out defines the impact of the ICT on freedom and security in the society as follows (Schermer and Wagemans, 2011): "Our modern-day Internet is an environment that allows for great freedom, but with freedom comes responsibility. If we want to keep the Internet an open, safe, and vibrant online environment, we must ensure that we take into account and protect the rights and interests of all members of society. The greatest challenge for Europe is to ensure the highest degree of (online) freedom for all."

The overall views on how important the ICT will be for the future society differs.

On one hand, Peter Drucker (2001) wrote: "All this suggests that the greatest changes are almost certainly still ahead of us. We can also be sure that the society of 2030 will be very different from that of today, and that it will bear little resemblance to that predicted by today's best-selling futurists. It will not be dominated or even shaped by information technology. IT will, of course, be important, but it will be only one of several important new technologies. The central feature of the next society, as of its predecessors, will be new institutions and new theories, ideologies and problems."

Similar is the conclusion by the National Intelligence Council in the "Global Trends 2025: "Many stress the role of technology in bringing about radical change and there is no question it has been a major driver. We—as others—have oftentimes underestimated its impact. However, over the past century, geopolitical rivalries and their consequences have been more significant causes of the multiple wars, collapse of empires, and rise of new powers than technology alone."

On the other hand Yochai Benkler (2006) writes: "Information, knowledge, and culture are central to human freedom and human development. How they are produced and exchanged in our society critically affects the way we see the state of the world as it is and might be; who decides these questions; and how we, as societies and polities, come to understand what can and ought to be done. For more than 150 years, modern complex democracies have depended in large measure on an industrial information economy for these basic functions. In the past decade and a half, we have begun to see a radical change in the organization of information production. Enabled by technological change, we are beginning to see a series of economic, social, and cultural adaptations that make possible a radical transformation of how we make the information environment we occupy as autonomous individuals, citizens, and members of cultural and social groups. It seems passe´ today to speak of "the Internet revolution." In some academic circles, it is positively naıve. But it should not be. The change brought about by the networked information environment is deep. It is structural. It goes to the very foundations of how liberal markets and liberal democracies have coevolved for almost two centuries."

Brynjolfsson and McAfee (2011) claim that we are yet to witness the most important impacts that the ICT will have on humanity. In the Moore's law they see its essential message – exponential growth – and compare the growth of the impact to the famous reward the inventor of chess was supposed to be given by the emperor of India. One grain of rice on the first square of the chessboard, two on the second four for the third … and so on.

Brynjolfsson and McAfee claim that we are still in the first half of the board where the reward, 2^{31} grains of rice, was quite manageable. But "as the technology moves into the second half of the chessboard, each successive doubling in power will increase the number of applications where it can affect work and employment." They pinpoint the source of this impact: "When businesses are based on bits instead of atoms, then each new product adds to the set of building blocks available to the next entrepreneur instead of depleting the stock of ideas the way minerals or farmlands are depleted in the physical world."

4 THE EXPECTATIONS OF FUTURE IMPACTS

The expectations of impact of ICT revolution on society are two-fold. The thesis is that it empowers the individual, the antithesis, that it empowers the state. We will propose the answer that both may be correct.

The thesis that technology empowers and brings freedom to individuals is called by Morozov and others "**the techno romantic view**". It is based on the assumption that ICT lowers the lowers transaction costs for people and that democratic availability of once exclusive technology puts relatively larger power into the hands of the people vis-à-vis the state. Democratic movements in Modova, Iran, Ukraine, Tunisia, Egypt, the tea party movement (USA), the OWS and the green movement (DE) are given as an example.

The alternative view is "**the techno-realistic view"**. The thinking is that ICT also lowers

transaction costs for governments. As the lives of the citizens and the work of the NGOs are increasingly digitized, it is easier to gather information, censor, and them crack-down. The internet architecture has a few spots that can be controlled if needed. As evidence, Cuba, China, Belarus, Iran and, recently, the NSA, are cited as an example.

The "**techno pragmatic"** view states that it depends. In democratic societies it empowers the citizens and NGOs while in totalitarian societies it empowers the state.

Regardless of the view, according to Scott's institutional theory institutions are social structures (wider term than organizations). They have the following basic aspects:

- regulative (setting and enforcement of rules),
- normative (values and social norms),
- cultural (interpretation of reality).

Because they are based on culture and values, institutions are very stable but they do change. Change of institutions (according to Tolbert and Zucker, 1983) is shown schematically in Figure below:

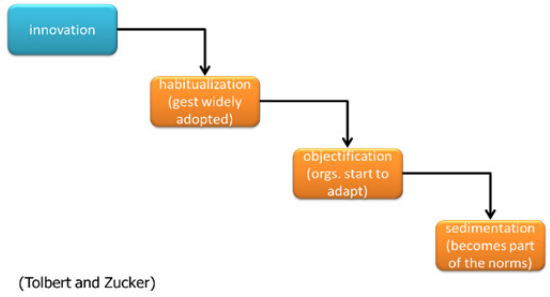


Figure 1: Cascade of change.

Historically it took centuries that the innovation of cheap paper lead to its wide adoption (as for books and newspapers). Businesses and individuals adapted accordingly. It took a few violent revolutions, however, some 300 years after the innovation of paper that the normative context (democracy, market economy) prevailed.

With the digital ICT much of the habitualization and objectification happened already. However, we are currently struggling even with the simple normative changes (e.g. copyright). Major changes to how society is organized are still to come. The study of this should be a major topic of the emerging internet science.

5 CONCLUSIONS

With respect to the impact of ICT on society, the author is convinced that Peter Drucker, the National Intelligence Council and others are wrong and that the cyber-utopians are right. The changes in the ICT will be one of the most important drivers of change of the future world.

To put an argument very shortly: for any cell of human society to work, its members must communicate. This holds true for families, businesses, local communities, states and countries as well as the globe as a whole. A change in communication possibilities influences these cells of society. The deeper the change, the larger the impact on societies.

The world has evolved into what it is – politically, economically, militarily, and culturally – around a material economy and analogue communication. Location mattered, borders existed. There is no such thing as distance and border in the digital communication world. Today, the digital economy is a small, but increasing part of the whole. What can be digitized – knowledge and meaning – constitute a growing share of the added value of products and services.

There are problems as to where and how to tax digital content, where and how should the digital nomads – people who work at a distance digitally – contribute to their social security, what "national" law applies for things that are somewhere in the internet, somewhere in the cloud; how to manage intellectual property rights for the kind of property where giving a copy to someone does not mean the giver is deprived of her own copy etc. Societal structures of the analogue world are not fit for the digital world any more.

The material economy of the 19th century largely took place inside a country. That county was a natural unit for rules, regulations and laws. The digital space is global. It will call for global rules and regulation. In the absence of which the actors that *can* play globally will rise in influence. This includes corporations as well as individuals. Which leaves us with the question of fair competition among and between those.

The author believes this change will be profound and will result in reinvention of all institutions of society. The process already started (as per Tolbert and Zucker) where citizens are free and where businesses must adapt in order to be competitive. The process is stalled where powerful institutions and legal frameworks are blocking change and leading to rigidities in society. In the next years and decades, much will depend on how change will happen in these areas – in highly governed and regulated areas of society that include public services, rule of law, education etc.

5.1 Towards Internet Science

Knowledge about this is needed on all levels of future thinking – from planning and strategies to forecasting and visions. A science that would provide an ontological, epistemological and methodological base is badly missing. It is a science that combines technologists with a clear understanding of where technology push and technology opportunity will be coming form, social scientists with an understanding of societal mechanisms, and humanities with a deeper understanding of the human being.

Internet science could fill-in this gap – if it becomes to computer science what urban planning and logistics are to civil engineering. Not developing the underlying technology on how to build the infrastructure but taking technological infrastructure for granted and studying how it can be used to improve the lives of people, create new businesses and interesting new work opportunities. The EINS network of excellence is making an important step in this direction.

ACKNOWLEDGEMENT

The work presented is supported, in part, by the The Network of Excellence in Internet Science funded by the European Commission.

REFERENCES

Aho, E. (2006) Creating an Innovative Europe, Report of the Independent Expert Group on R&D and Innovation, http://ec.europa.eu/invest-in research/pdf/download_en/aho_report.pdf

Benkler, Yochai (2006), The wealth of networks : how social production transforms markets and freedom (1st ed.), New Haven, Conn: Yale University Press, p. 528, ISBN 0-300-11056-1

Brynjolfsson, Erik and Andrew McAfee (2011). Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy, Digital Frontier Press.

Drucker, Peter (2001). The Next Society, The Economist, November 2001.

European Commission (2009). Lisbon Strategy for Growth and Jobs, http://ec.europa.eu/archives/growthandjobs_2009/

European Commission (2009a). Future Internet 2020, Visions of an Industry Expert Group, http://www.future-internet.eu/fileadmin/documents/reports/FI_Panel_Report_v3.1_Final.pdf

European Commission (2009b). THE WORLD IN 2025, Rising Asia and socio-ecological transition, http://ec.europa.eu/research/social-sciences/pdf/the-world-in-2025-report_en.pdf

European Commission (2010). Digital agenda for Europe 2010-2020 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0245:FIN:EN:PDF

European Council (2009). Future of ICT research, innovation and infrastructures - Council conclusions http://register.consilium.europa.eu/pdf/en/09/st17/st17190.en09.pdf

European Internet Foundation (2009). The Digital World in 2025, Indicators for European Action. http://www.EIFonline.org/

European Union (1994). Bangemann Report, Europe and the Global Information Society, http://www.umic.pt/images/stories/publicacoes200801/raport_Bangemanna_1994.pdf

Forge, S., Blackman, C., Popper, R., Guevera, K., Srivastava, L., Brown, I., Cave, J., and Miles, I. (2010), Towards a Future Internet: Delphi Survey Round 1 Results, A report to EC DG Information Society and Media, Oxford Internet Institute and SCF Associates.

Gartner (2010). Top 10 Strategic Technologies for 2011, http://www.gartner.com/it/page.jsp?id=1454221

Kurzweil, Rey (1998). The Age of Spiritual Machines: When Computers Exceed Human Intelligence, Penguin 0-14-028202-5.

Morozov, Evgeny (2010). The Net Delusion: The Dark Side of Internet Freedom.

National Intelligence Council (2009). Global Trends 2025: A Transformed World - The National Intelligence Council's 2025 Project, http://www.cfr.org/world/global-trends-2025-transformed-world--national-intelligence-councils-2025-project/p17826

Paradiso Project (2011). http://paradiso-fp7.eu/

Pink, Daniel H. (2006). A Whole New Mind: Why Right-Brainers Will Rule the Future, Penguin Books

Reflection Group (2010), Project Europe 2030, Challenges and Opportunities, http://www.consilium.europa.eu/uedocs/cms_data/librairie/PDF/QC3210249ENC.pdf

Rifkin, Jeremy (2009). The Empathic Civilization, Polity Press, Cambridge.

Schermer, Bart and Ton Wagemans (2011). Freedom in the Days of the Internet, Centre for European Studies, Brussels.

Tolbert, Pamela S. and Lynne G. Zucker (1983), Institutional sources of change in the formal structure organizations: The diffusion of civil service reform, *Administrative Science Quarterly*, vol. 30, 1983.

Turk, Žiga (2009). Communication Revolutions Change Civilizations: What to Expect from the Current One?, keynote paper, *6th International WEBIST Conference*, http://www.webist.org/WEBIST2010/Keynote_Speakers.htm

Zittrain, J. (2008) The Future of the Internet: And How to Stop it, Yale University Press, New Haven, CT, http://futureoftheinternet.org/download

Utilising Social Media Technology to Raise Brand Awareness in Higher Education

Baomin Qi[1] and Lindsey Mackie[2]
[1]*Bolton Business School, University of Bolton, Deane Road, Bolton, U.K.*
[2]*Career and Employability, University of Salford, University House, Salford, U.K.*
BQ2@bolton.ac.uk, l.mackie@salford.ac.uk

Keywords: Social Media Technology, Brand Awareness, UK Higher Education.

Abstract: Marketing and more specifically raising brand awareness is vital for any successful organisation and its importance is demonstrated across all industry sectors. The widely adopted social media (SM) technology is believed to bring opportunities for brand awareness; and many universities worldwide have already broached the concept of using SM to communicate with potential students. However, the social media arena is deemed by many HEIs as frivolous, and the significance of SM is yet widely realised. This research will look at the power behind social media technology; the main focus is to explore whether the utilisation of SM technology could increase the brand awareness. In order to achieve this research aim exhaustive literature search was conducted and the available literature is systematically reviewed. The research found that although SM platforms have been exploited by most of institutions, there is little evidence to suggest that higher education is currently using social media effectively to engage with students, increase brand awareness and enhance image.

1 INTRODUCTION

The term social media is often associated with the term Web 2.0 which includes social networking sites (SNS) such as Facebook, Blogging sites e.g. Blogger.com and micro-blogging sites e.g. Twitter. Within this umbrella also includes Video (e.g. YouTube), Audio (e.g. iTunes and Podcast.net) and live-casting (e.g. BlogTalkRadio and Live 365) and the virtual worlds (e.g. Second Life). This clearly outlines the complexity surrounding the term social media, however the one commonality between all of the above areas is that they all focus on building online communities within which people are interacting and influencing each other.

Universities worldwide have adopted social media to an extent and whilst some have successfully recruited students to their programmes using social media marketing (Reuben, 2008) the majority use it mostly to stay in touch with graduates and current students. Although there are numerous articles written regarding the power of social media marketing, very little is available in relation to the public sector and in particular the Higher Education sector. Also it has been identified that marketers in the public and private sectors have extremely differing viewpoints on the relevancy and importance of social media.

The primary purpose of this research is to examine the effect of social media presence on brand awareness in higher education. With this in mind, a literature search was conducted using Emerald Insight and ProQuest journals in the areas of social media marketing and higher education. Three key themes emerged from the literature search, which were: use of social media for marketing; use of social media to raise brand awareness (and image and reputation) and use of social media for learning. Key research in these areas of interest will be discussed in more detail in the following sections.

2 OVERVIEW OF SOCIAL MEDIA ADOPTION IN HE

Studies of US universities have shown that many are now using SNS to conduct marketing targeted to the individual student's needs (Wandel, 2008; Mattson and Barnes, 2009). The study by Mattson and Barnes (2009) was concerned with how admissions

departments are using social media to connect with students in order to attract them to their programmes. They also used social media to research prospective students. Whilst this report demonstrated some success in engaging with students this was with the purpose of recruitment and it was impossible to say whether brand awareness was raised though its use. However they also discovered that only half of the universities researched were using social media to monitor any conversations about them online.

Reuben's (2008) research also showed that universities across America are using social media in their marketing and recruitment efforts as well as to communicate with current students and maintain contact with alumni. These studies agreed that institutions need to better understand what their students use SNS for by learning the "rules of engagement". This is confirmed by the findings of Boon and Sinclair (2009); Hemmi et al. (2009) and Madge et al. (2009) who all raised concerns about appropriate channels of communication between university and student and said that social media should be used with caution in education. However, these studies were concerned with the interaction amongst students and academics when used as a platform to support learning and engagement, rather than the use of social media to raise brand awareness. Concerns were also raised about the loss of control as institutions felt that having a presence on social media could leave them open to abuse and negative comments. Other studies have shown that being able to respond to online negativity enables brands to start a conversation with their communities and change perceptions of their organisation (Sharp, 2011).

Minocha (2009) looked at how SNS were being used to support student learning at 18 Higher Education Institutions in the UK. This highlighted a number of benefits such as improved team building and peer group support. It is generally considered acceptable by students to communicate with their classmates in this way to enhance their learning experience but it was found that most students did not welcome the idea of communicating with their lecturers on their personal social spaces. The study also raised concerns that the institutions would have little control over the public social media sites The Open University has a Facebook application called 'My OU story' which allows prospective students to interact with current students (Redecker, 2009). This offers the benefit that students can find out more about their prospective university and intended course to support their decision about where to study. Current students are giving positive messages about their experience at the OU and this enhances the online reputation of the university and also their brand image. The University of Westminster has a similar social networking forum aimed at staff and students (Oriadni and Saunders, 2008) which enables students to interact with their lecturers giving a positive learning experience. Smith et al (2007) agree that SNS can be used by universities to generate word of mouth (WOM) publicity to promote their institution and its achievements. If this is true, then it is reasonable to assume that promoting the successes of an institution should increase brand awareness and image.

Whilst not concerned specifically with social media, the findings of Mazzarol and Soutar (2012) highlight how institutions of higher education are tacking the issue of branding. It was found that branding was not approached in a strategic manner by most universities and was a combination of online and offline resources, including brochures, prospectuses, location and rankings. They also found that universities often have several brands within the institution, for example a sub-brand for each separate school, faculty, research department, etc. Whilst this might not seem to be an issue, the information about how students use social media would suggest they might struggle to identify which brand they should be engaging with when presented with a choice of many which can effectively weaken the overall brand.

3 SM USAGE FOR LEARNING

The majority of social networking sites users are of university age and it therefore makes sense that universities would want to capitalise on this additional communication channel. However, some studies (Boon and Sinclair, 2009; and Madge et al, 2009) reported that students have reacted unfavourably to the blurring of boundaries between formal relations with tutors in class and informal discussions on SNS. Many students consider their interactions on social media platforms as private (between their friends and family) and see this as a social medium rather than one to engage with official bodies. Whilst most people will like a number of business pages on Facebook, their engagement with these pages is limited and Kalpana and Anandan (2013) found that any content posted needs to be designed to suit the target audience. Their study examined ten higher education institutions in India and they concluded that

organisations need to aim to connect with their community by finding out about their lives and becoming part of that, rather than simply posting marketing messages. Their findings suggested that the type of content was also important with messages containing pictures and links far more likely to attract attention than those that were text only. According to Facebook (2013) status updates containing pictures get 120% more attention than those without. In addition, shorter messages (less than 250 characters) are felt to gain more interaction. Similar conclusions have been reached from commercial organisations with regard to the use of Facebook for engaging with customers.

Isacsson and Gretzel (2011) developed a collaborative project between a university in Texas, USA and another in Finland. They set up a Facebook group where students from both universities could collaborate whilst learning about sustainable tourism. Whilst the lecturers involved put a lot of time into setting up the social media platform and some students did find it useful to collaborate in an informal environment, other students were uncomfortable about communicating with people that they did not really know. Although the project was considered a success it highlighted that participants in different countries have different views on how to engage using social media. The authors concluded that they had to coerce students to participate in discussions and encourage them to post on the message boards, which confirmed a belief that when students connect with their institution in this manner they see it as a means of accessing information, rather than a medium for two way communications.

Shaw (2013) found that whilst the majority of students use social media more than once a day, they are either unaware of their university's presence or do not consider the information trustworthy or relevant to their needs. It was felt that universities need to be more responsive to their students' requirements. Lilley et al (2012) found that 94% of the students they surveyed consider Facebook important for keeping in touch with friends and family and socialising, whilst over half said they did not consider it important for receiving advertising messages and many of those questioned said that they often ignored such messages anyway.

In the main, young people appear to be using social media to connect with friends and to socialise, rather than be marketed to. In terms of university students, they do use social media to help settle in to life on campus, make new friends and interact with their classmates (Wandel, 2008). Whilst they might use it in this way, most seemed less inclined to engage with academic staff and might not interact with the official university presence on social media.

4 USE OF SM FOR DIRECT MARKETING

Organisations develop strategic marketing plans to promote and sell their services and more recently this has included social media marketing. This is a relatively new concept, meaning to gain attention or website traffic through the use of social media (Saravanakumar and SuganthaLakshmi, 2012). It has developed since the arrival of social networking sites (SNS), particularly Facebook in 2004 and Twitter in 2006 (Boyd and Ellison, 2007; Newson, 2009). With the advent of social networking organisations have also come to realise that whilst reputations can be built using social media, they can quickly be devastated by negative feedback.

Whilst some researchers are critical of the use of SNS for direct marketing (Dury, 2008; Odden, 2009), it has been proven to be important in raising brand awareness and image. Dury (2008) argues that whilst some believe marketing has no place in social media because it is a business activity rather than a social one, the face of marketing has now changed and whilst social media marketing is felt by many to be unacceptable, marketers must rise to the challenge. Odden (2009) argues that people join SNS to interact with others who have shared interests, not to be bombarded with advertisements. However, the commercial sector has successfully used SNS to sell products to a wide audience and Saravanakumar and SuganthaLakshmi (2012) described how several retailers have used social media to listen to their fan base in order to design offers or products which match their clients' needs. Gap, for example, invited customers to debate their new logo online and after reviewing feedback decided to revert to their original logo at their fans' request. Other sectors are following suit with social media at a slower pace with varying degrees of success.

By its very nature social media marketing centres its target audience during the design of marketing campaigns (Smith and Zook, 2011). This begins with analysing data on the intended audience; this information informs the design and delivery of the marketing campaign to ensure it is highly relevant to the target audience. In a study of young female students, Logan et al (2012) discovered that

advertising via SNS is more effective when it is entertaining rather than informative and Wang (2012) found that brands targeting younger audiences had higher engagement on SNS. Blanchard (2011) believes that campaigns which encourage engagement and raise awareness are better received on SNS than those which are simply trying to sell products. Overall, this demonstrates that whilst direct marketing may not be appreciated by the younger generation, across the typical age range for university students, social media does hold potential for engagement and interaction which can increase brand awareness and image.

Social media can also act as a communications tool which provides instant feedback (Blanchard, 2011). Evans (2011) stresses the importance of not only listening to customers but also taking the opportunity to respond to any negative comments to resolve issues and Clapperton (2009) gives examples of organisations that turned around relationships with disgruntled customers who became brand advocates.

5 USE OF SM TO RAISE BRAND AWARENESS

Commercial organisations quickly adopted SNS to reach their customer base and many have used social media to raise brand awareness and increase engagement with their customers. Li and Bernoff (2008) believe the key to using social media successfully lies in building a strong relationship and raising brand awareness amongst customers. Booth and Matic (2011) discovered that socialising a brand meant organisations were able to engage one on one with their customers in order to build communities and brand advocates. Consumers are using social media to make themselves heard by the organisations they want to do business with and Saravanakumar and SuganthaLakshmi (2012) agree that social media marketing has been shown to be particularly successful in raising awareness. Mangold and Faulds (2009) said that as social media enables companies to communicate with their customers and for customers to talk to one another this enables communities to be built up around a brand. Whilst this can help raise brand awareness and reputation if the discussions are spreading positive messages, social media can damage an organisation's brand if participants are making negative comments or giving poor feedback. Aula (2010) gives three examples of organisations who suffered due to messages which were quickly spread via social media and ultimately picked up on by mass media so that the stories were publicised worldwide on television and in tabloid newspapers. In the case of United Airlines an unhappy customer posted a video of himself singing a song entitled "United Breaks Guitars" this quickly went viral and the video received millions of views on YouTube. The negative publicity resulted in loss of income for the organisation. This demonstrates the potential threat to both the reputation and business of companies which can quickly be damaged in the social media arena.

Sharp (2011) demonstrated how an organisation can prevent damage to their online reputation by listening to the conversations that are taking place and reacting appropriately. Responding to comments made on social media meant that they were able to demonstrate their excellent customer service by resolving issues quickly.

Bruhn et al. (2012) studied three industry sectors to see how social media compared to traditional media in terms of raising brand awareness and whilst traditional media was felt to be more influential in shaping brand awareness, social media was more effective in enhancing image. Sharp (2011) noted how a healthcare clinic in Cleveland was able to successfully use social media to promote its services and in doing so to raise awareness by the positive WOM spread by clients. Some sectors felt using SNS was less appropriate and banks in particular shied away from it as they felt customers did not wish to interact with them in this way, although they did not rule out its potential for the future (Mitic and Kapoulas, 2012).

Wang (2012) gives a good example of a Singapore university adopting social media to increase engagement with their students using a competition as the incentive. Students were asked to like the page and post a comment (thus increasing likes and engagement). They then had to share this comment on their own page and encourage friends to like it, further increasing engagement and awareness. The student who received the most "likes" won an iPad.

6 CONCLUSIONS

The initial literature review has shown that SNS such as Twitter and Facebook are being used by organisations to raise brand awareness and improve their image with customers. The area of social media is somewhat broad and previous studies have not

delved into higher education specifically with regard to how it is being used to raise brand awareness.

A new study by Schultz and Peltier (2013) suggests that social media so far has been used to market to existing customers who are already engaged with a brand and the findings imply that it is now time to examine ways in which social media can be used to build engagement with new customers rather than using it to increase sales in the short term. Although its use in the commercial sector has been fairly well documented and it has been recognised as a vital tool in marketing strategy, this concurs with the findings that it is usually in the short term.

The authors found little evidence to suggest that higher education is currently using social media effectively to engage with students, increase brand awareness and enhance image. Mazzarol and Soutar (2012) stress the importance of brand management for universities and believe that in this highly competitive time those that have an "established reputation" are most likely to survive. They do, however, conclude that this is an area which is not yet managed effectively in UK institutions and the author feels this is an area that merits further research. Frydenberg, (2013) believes that social media "holds promise... for engaging students..." whilst Kalpana and Anandan (2013) found that engagement with students can be increased if the content considers the needs of the target audience. This suggests that universities need to discover exactly what students want to see posted on social media by their institutions.

The next stage of this research project will focus on the strategic development of using social media technology to raise university's brand awareness. Case study approach will be adopted and a range of case studies will be carefully chosen.

REFERENCES

Aula, P. (2010). Social media, reputation risk and ambient publicity management. Strategy & Leadership [Online] 38 (6), pp.43-49. Available from: <http://dx.doi.org/10.1108/10878571011088069> [Accessed 23 April 2013]

Boon, S. and Sinclair, C. (2009) A world I don't inhabit: disquiet and identity in Second Life and Facebook. *Educational Media International* [Online]. *46*(2), pp 99-110. Available from: <http://www.informaworld.com/smpp/content~db=all~content=a911901158> [Accessed 12 February 2013].

Booth, N. and Matic, J.A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. *Corporate Communications* [Online]16 (3) pp 184-191. Available from: <http://search.proquest.com/docview/888255196?accountid=9653> (accessed 21 March 2013).

Boyd, D.M. and Ellison, N.B. (2007) Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* [Online]. *13*(1) article 11. Available from: <http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html> [Accessed 3 March 2013].

Bruhn, M., Schoenmueller, V. and Schäfer, D.B. (2012) Are social media replacing traditional media in terms of brand equity creation?. *Management Research Review* 35(9) pp.770 – 790. [Online]. Available from: <http://dx.doi.org/10.1108/01409171211255948> [Accessed 22 March 2013]

Cheater, M. (2008) How Higher Ed is Using Facebook Pages [Online]. Available from: <http://www.academicagroup.com/node/4598> [Accessed 11 March 2013].

Clapperton, G. (27 Oct 2009) *This is Social Media: Tweet, Blog, Link and Post Your Way to Business Success.* Chichester, West Sussex: Capstone Publishing (Wiley)

Dury, G. (2008) Opinion piece: Social media: Should marketers engage and how can it be done effectively. *Journal of Direct, Data and Digital Marketing Practice* [Online]. 9 (3) pp. 274-277. Available from: <http://www.palgrave-journals.com/dddmp/journal/v9/n3/full/4350096a.html > [Accessed 12 March 2013].

Evans, L. (2010) *Social Media Marketing: Strategies for Engaging in Facebook, Twitter & Other Social Media: Promoting Your Company...* USA: Que Publishing

Facebook (2013) *Facebook for Business. Build your Page*. [online]. Available from: <https://www.facebook.com/business/build> [Accessed 28 April 2013]

Frydenberg, M. (2013) Creating a collaborative learning community in the CIS Sandbox. *Interactive Technology and Smart Education.* [Online] 10(1) pp.49–62.

Hemmi, A., Bayne, S. and Land, R. (2009) The appropriation and repurposing of social technologies in higher education. *Journal of Computer Assisted Learning* [Online]. 25(1) pp. 19–30. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2729.2008.00306.x/full

Isacsson, A. and Gretzel, U. (2011) Facebook as an edutainment medium to engage students in sustainability and tourism. *Journal of Hospitality and Tourism Technology.* 2(1) pp. 81 – 90.

Kalpana, C. and Anandan, P. (2013) Role of content strategy in social media brand communities: a case of higher education institutes in India. *Journal of Product & Brand Management*. [Online]. 22(1) pp.40 – 51.

Lilley, S., Grodzinsky, F.S. and Gumbus, A. (2012) Revealing the commercialized and compliant Facebook user. *Journal of Information, Communication and Ethics in Society*. 10(2) pp.82 –

92 [Online]. Available from: <http://dx.doi.org/10.1108/14779961211226994> [Accessed 30 April 2013]

Logan, K., Bright, L.F. and Gangadharbatla, H. (2012) Facebook versus television: advertising value perceptions among females. *Journal of Research in Interactive Marketing* [Online] 6(3) pp.164 – 179. Available from: <http://dx.doi.org/10.1108/17505931211274651> [Accessed 20 April 2013]

Madge, C., Meek, J., Wellens, J. and Hooley, T. (2009) Facebook, social integration and informal learning at university: 'It is more for socialising and talking to friends about work than for actually doing work'. *Learning, Media and Technology* [Online]. 34(2) pp.141-155.

Mangold, W.G. and Faulds, D.J. (2009). Social Media: The new hybrid element of the promotion mix. *Business Horizons* [Online]. 52 pp. 357-365. Available from: <http://japanfashionshanghai.up.seesaa.net/image/japanfashionshanghai-2010-07-06T12:32:54-2.pdf> [Accessed 11 February 2013]

Mattson, E., and Barnes, N. G. (2009) *Social Media and College Admissions: The First Longitudinal Study* [Online]. Available from: <http://www.neacac.org/docs/uploads/files/amc2010/2010%20AMC%20Web%20Social%20Media%20and%20College%20Admissions%20The%20First%20Longitudinal%20Study.pdf> [Accessed 12 March 2013].

Mazzarol, T. and Soutar, G.N. (2012) Revisiting the global market for higher education. *Asia Pacific Journal of Marketing and Logistics* [Online] 24(5) pp.717 – 737. Available from: <http://dx.doi.org/10.1108/13555851211278079> [Accessed 18 March 2013]

Minocha, S. (2009) *A Study of the Effective Use of Social Software by Further and Higher Education in the UK to Support Student Learning and teaching* [Online] Available from: <http://kn.open.ac.uk/public/document.cfm?docid=12120> [Accessed 24 January 2011].

Mitic, M. & Kapoulas, A. (2012) "Understanding the role of social media in bank marketing", Marketing Intelligence & Planning, 30(7) pp.668 – 686. . [Online]. Available from: <http://dx.doi.org/10.1108/02634501211273797> [Accessed 22 March 2013]

Odden, L. (2009) How Direct is Social Media Marketing, Online Marketing Blog, 23 June [Online]. Available from: <http://www.toprankblog.com/2009/06/howdirect-is-social-media-marketing/> [Accessed 12 January 2011].

Oriadini, F. and Saunders, G. (2008) Social networking: Connect-ing students and staff. *Association for Learning Technology News* [Online]. *(*62) p.3. Available from: <http://www.alt.ac.uk/docs/alt-n_62_final.pdf> [Accessed 18 February 2013].

Redecker, C. (2009) Review of Learning 2.0 Practices: Study on the Impact of Web 2.0 Innovations on Education and Training in Europe [Online]. Available from: <http://ftp.jrc.es/EURdoc/JRC49108.pdf> [Accessed 17 March 2013].

Reuben, R. (2008) Survey about the use of social media in higher education [Online]. Available from: <http://www.docstoc.com/docs/20087333/The-Use-of-Social-Media-in-Higher-Education---A-Guide> [Accessed 28 March 2013].

Saravanakumar, M. and SuganthaLakshmi, T. (2012) Social Media Marketing. *Life Science Journal* [Online] 9(4) pp 4444-4451. Available from: <http://www.lifesciencesite.com/lsj/life0904/670_13061life0904_4444_4451.pdf> [Accessed 22 March 2013]

Schultz, D. and Peltier, J. (2013) Social Media's Slippery Slope: Challenges, Opportunities and Future Research Directions. *Journal of Research in Interactive Marketing*. 7(2) [online]. Available from: <http://www.emeraldinsight.com/journals.htm?issn=2040-7122&volume=7&issue=2&articleid=17087504&show=abstract> [Accessed 30 April 2013]Sharp, J. (2011). Brand awareness and engagement: A case study in healthcare social media. *Frontiers of health services management*. [Online] 28(2) pp.29-33 Available from: <http://search.proquest.com/docview/912463726?accountid=9653> [Accessed March 20, 2013]

Shaw, J. (2013) University recruitment: one fifth of students say social media doesn't work. The Guardian. 17 April 2013. [Online]. Available from: <http://www.guardian.co.uk/higher-education-network/blog/2013/apr/17/university-student-recruitment-social-media> [Accessed 30 April 2013]

Smith, T., Coyle, J. Lightfoot, E. and Scott, A. (2007) Reconsidering Models of Influence: The Relationship between Consumer Social Networks and Word-of-Mouth Effectiveness. *Journal of Advertising Research* [Online]. 47(4), pp. 387. Available from: <http://www.communityanalytics.com/Portals/0/Resource_Library/WOM%20in%20Journal%20of%20Advertising.pdf> [Accessed 4 April 2013].

Wandel, T. (2008) Colleges and Universities Want to be Your Friend: Communicating via Online Social Networking. *Planning for Higher Education* [Online]. *37* (1) pp.35-48.

Wang, H. (2012) Six P's of youth social media from a young consumer's perspective. *Young Consumers: Insight and Ideas for Responsible Marketers* [Online] 13(3) pp.303 – 317. Available from: <http://dx.doi.org/10.1108/17473611211261674> [Acessed 12 March 2013]

MOBILE INFORMATION SYSTEMS

FULL PAPERS

Cascading Information for Public Transport Assistance

Christian Samsel[1], Shirley Beul-Leusmann[2], Maximilian Wiederhold[1],
Karl-Heinz Krempels[1], Martina Ziefle[2] and Eva-Maria Jakobs[2]
[1]*Information Systems, RWTH Aachen University, Aachen, Germany*
[2]*Human-Computer Interaction Center, RWTH Aachen University, Aachen, Germany*
{*samsel, wiederhold, krempels*}*@dbis.rwth-aachen.de*
{*beul-leusmann, ziefle*}*@comm.rwth-aachen.de, e.m.jakobs@tk.rwth-aachen.de*

Keywords: Context-Awareness, Gamification, Travel Information, Mobile Applications, Usability.

Abstract: Over the last years, public transport has become both more prominent and more diverse. Because of the complex structure of today's public transport networks, an electronic guidance is effectively required. Usually different transport modalities and service providers offer their own application to which the traveler has to adapt after changing between services. Additionally a current trend in mobile applications is the customization of GUI elements which leads to appealing looks but usually also to cluttered presentation of information. Both these problems cause a high cognitive stress on the traveler using the mobile application, especially while conducting other activities at the same time. Our approach to mitigate these issues is to create a mobile application applying the Gamification principle *Cascading Information Theory* to simplify the usage and additionally to use a back-end which allows to integrate data from various services hereby unifiying the presentation. A prototype of the application was evaluated in an initial user test for comparing our approach to the most popular mobile travel application in Germany.

1 INTRODUCTION

Global trends like urbanization, climate change, and peak oil enforce and accelerate the change of people's mobility; the focus shifts back from individual transport to public transport. Traditional public transport networks, like subways and bus networks, have grown in size and new forms of transport systems like carsharing have been introduced. Consequently, the number of potential itineraries from one point to another raises, each with specific features, e.g., different prices and comfort levels. To use public transport efficiently, meaning to decide on the most suitable itinerary and to follow it closely, the traveler requires effective electronic assistance. To do so a large number of more or less appropriate travel information systems have been established. Such travel information vary from traditional web platforms used for plane or train ticket booking to modern mobile applications for sharing rides.

1.1 Motivation

From our continuous work on travel information, indoor navigation, diversity, and mobile applications we recognized that most current mobile applications for public transport are flawed in terms of usability under cognitive stress. Current applications (see Section 2.1) shine with a fancy Graphical User Interface (GUI), many configuration options and a high amount of information. Such applications perform well in situations where the traveler can direct all of her cognitive capacity to the application, but fall back in terms of usability in stressful situations. Stress can be introduced by noise, time pressure, unknown surroundings, or required multitasking (e.g., walking and using a device at the same time).

To improve the usability under stress we wanted to create a concept and a prototype which simplifies mobile travel assistance compared to already existing applications. While creating a concept based on our experience in indoor navigation (Heiniz et al., 2012) we realized that the initial GUI layout and interaction pattern is strongly related to archetypical Massively Multiplayer Online Role-Playing Games (MMORPGs) like World of Warcraft (Corneliussen and Rettberg, 2008). World of Warcraft, for example, uses non-intrusive, still useful indicators/descriptions for the next waypoint in a quest. The user has access to exactly the required information she needs to ful-

fill the given task (e.g. travel to a specific location) at one glance. This principle of providing exactly the required information in the current context is called Cascading Information Theory. We wanted to transfer this principle to a real world application.

1.2 Cascading Information Theory

In his 2010 blog entry[1], Erick Schonfeld issued his company's "playdeck" of game dynamics terms which included most if not all of the currently employed elements of gamification. Gamification is a term which was coined by Nick Pelling in the early 2000's. Since then many different interpretations of the term have surfaced, making its exact meaning hard to conceive. Recently, there have been efforts by several researchers to reach a mutually agreed on definition. One of the more elaborate and most cited versions originates with (Deterding et al., 2011). They define gamification as: "The use of game design elements characteristic for games in non-game contexts." Elements of games are components or interaction patterns that, in combination, create the game experience. By abstracting these elements from their game implementation and employing them in another context (e.g. business software etc.) the developer makes use of people's play instinct. If a user is familiar with the ported game element there is a high probability that she will associate it with its original purpose and therefore be able to correctly utilize it.

Along with well known Gamification components like achievements, Schonfeld lists the principle of *Cascading Information Theory*. Cascading Information Theory suggests to unveil information about the game in as small amounts as possible to ensure the user's focus exactly on the desired objective. Thereby, confusion and misdirection of players by providing excess information is prevented and each iteration of new data can be applied directly.

The remainder of this paper is structured as follows: In Section 2, existing applications and related research are discussed. Section 3 describes our approach to the problem on an abstract level, whereas Section 4 demonstrates the actual prototype implementation. The evaluation is presented in Section 5 and Section 6 concludes the paper.

2 RELATED WORK

This section lists popular related applications and recent research.

[1] http://techcrunch.com/2010/08/25/scvngr-game-mechanics/

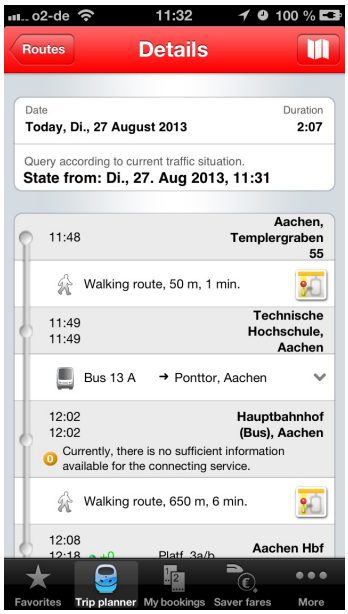


Figure 1: The DB Navigator application includes all trip data plus additional texts and various buttons in one view which impedes traveler orientation.

2.1 Public Transport Applications

The following part exhibits commonly used and high rated mobile applications for public transportation in Germany, not only explaining the benefits but also elaborating on flaws and drawbacks of the individual application, in order to convey the motivation for a new interface design. We considered only European applications because travel information applications are only meaningful applicable in the area for which they are created.

DB Navigator

Figure 1 shows a screenshot of the most popular (Statista GmbH, 2012) application for public transport in Germany, the DB Navigator. The DB Navigator is developed by HaCon AG on behalf of Deutsche Bahn. Therefore the assumption that its prevalence is due to the fact that its publisher is the main German provider of inter-regional public transportation suggests itself. The application also features fully compatible intermodal trip planning and supports real-time information about delays and arrival times. In addition to these functions there is also support for rudimentary walking directions and direct online purchase of tickets.

Despite its vast amount of features, the decent appearance, and the possibility of providing the company's own data the application has a few drawbacks.

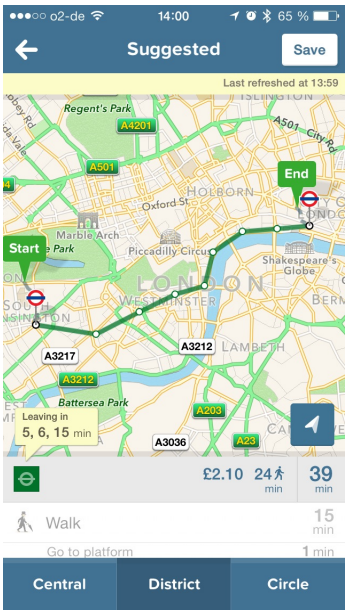


Figure 2: The London Citymapper is a very advanced application which is specifically adjusted to the needs of the London inhabitants.

The trip view shows the complete trip information in one tabular environment which requires the traveler to scroll vertically if the trip involves more than one change. To get walking directions the traveler has to press the button which resembles a map. The high number of GUI elements, the inconsistent coloring and required interaction complicate the interpretation and filtering of the needed information for the traveler and might delay the usage especially in a high stress situation.

Citymapper

The Citymapper application (London version tested) is well-known in England. The application includes a broad range of features such as generating intermodal trip plans using real-time information. Having selected the departure and destination locations the application suggests a variety of different possibilities to reach the desired spot. In addition to the intermodal route suggestions the user is presented with variations like routes only involving buses or rain safe ones and information about the current context such as the weather at the destination. The route detail view consists of a large map view displaying the selected trip itinerary and a table below it listing the individual parts of the journey (see Figure 2).

The Citymapper application addresses many of the crucial, common issues of public transportation and is appealing. However, it is specifically designed to match the London network transportation. This is a major drawback, adapting to a different application based on the whereabouts requires cognitive resources. This issue is exacerbated by the quantity of features which might be difficult to comprehend in a short amount of time.

2.2 Existing Research Work

This section presents recent work in the area of public transportation and mobile navigation systems and aims to give a short insight into the current state of research.

An attempt to create a multimodal routing service with directions for pedestrians is made in (Yu and Lu, 2012). In addition to evaluating travel modes by the criteria of distance, time, fare, and transfer, they use dynamic information about real-time and historic traffic data to enable accurate estimations. The computation of transit routes is aided by a high priority mode expansion strategy which incrementally generates parts of the most suited path. Similarly, the sophisticated algorithm developed by (Zheng and Xie, 2011) allows for accurate generic and personalized recommendations solely based on the analysis of user GPS traces. By correlating the travel behavior of users with their locations they are able to estimate user interests.

(Baus et al., 2002) introduce their concept of a pedestrian navigation system with a hybrid solution titled Project REAL. Already accurately predicting the evolution of personal navigation systems ("Personal navigation systems that extend beyond today's use in cars will play a major role in the future.") they present their approach featuring a combination of indoor and outdoor navigation. The system is capable of automatically adapting to location changes and presenting directions to the traveler on a variety of different devices.

Since the primary instruments for outdoor navigation are geolocation services, accomplishing indoor navigation is an incomparably more difficult task because exact positions are hard to determine. Nevertheless, considering the topic is crucial for any attempt at implementing a complete navigation system. In (Rehrl et al., 2005) a solution for the guidance of pedestrians inside transport interchange buildings is proposed which relies on generating a hierarchical model of the complex and providing instructions augmented with map views. (Chowaw-Liebman et al., 2010; Heiniz et al., 2012) present a turn-by-turn approach to manage navigation inside large, complicated buildings using landmark representations for orientation rather than street names. The work is

based on the idea that pedestrians mostly rely on landmarks as navigation cue (May et al., 2003).

The TransitGenie project (Biagioni et al., 2009) aims at implementing a public transportation navigator with a transit tracking service which consists of an activity classification algorithm to determine user context as well as trajectory matching to enhance vehicle tracking. In their later work, (Biagioni et al., 2011) further pursue the automatic real-time tracking of transit traffic and the computation and prediction of arrival times.

The UbiBus project started an attempt to create an ubiquitous system for public transport assistance using context information (Vieira et al., 2011). The authors incorporated access to an intelligent transportation system application in buses as well as web services and smart phone apps, so that passengers are enabled to receive information about the progress of their bus. (Pessemier et al., 2013) proposed a framework which is capable of extracting high-level information about user context. Initially the system recognizes and collects very basic data to create a foundation. Complex user activities are then broken down into smaller parts and connected via conditional relationships which are afterwards matched with the previously recorded basic contexts. If a complex context is identified this way, an application is capable of requesting detailed activity recommendations from a central server using a four-part scoring model. (Christoph et al., 2010) proposed an integrated context-detection service architecture which combines information from different approaches.

(Keller et al., 2011) observed that creating an ideal representation of public transit trip information on a mobile device is a difficult task. Using paper prototypes, the authors compare different approaches to display intermodal travel chains.

3 APPROACH

While working on public transportation information systems we identified parallels between real world navigation and navigation in, e.g., MMORPGs. Navigation elements in games are usually lean and unobtrusive as to not distract the player from her main task, but still are easy to comprehend. In a real world environment full of distractions, noise, and confusion the traveler's concentration can rarely be focused entirely on operating her mobile device. When traveling on a bus, entering a rail station, or even just walking towards a bus stop the traveler needs to be capable of checking her itinerary as quickly as possible. Therefore, an assistant application is required to be uncomplicated and easy to use. In order to minimize the amount of cognitive resources needed by the navigator, the user interface needs to be plain and responsive, similar to navigation elements in MMORPGs.

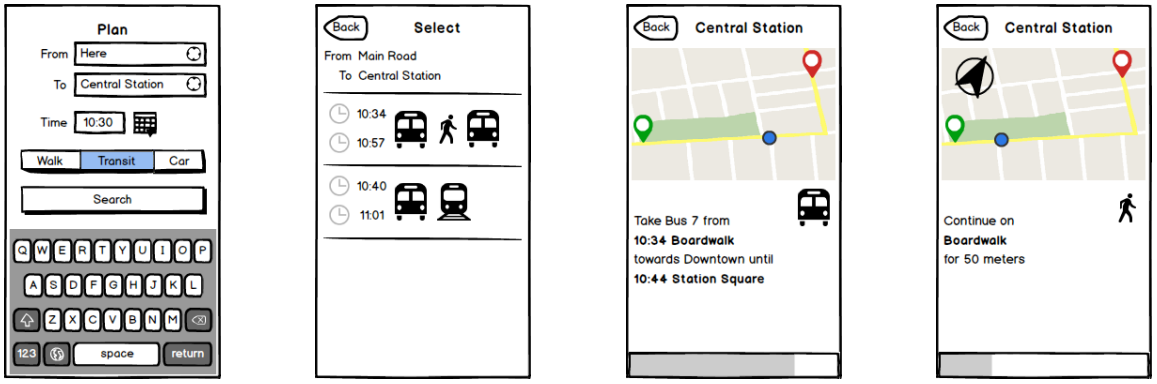
Following the general principle of Cascading Information we divided and analyzed information presented by mobile travel information applications regarding necessity. We only considered essential information for inclusion in our GUI concept. For every piece of essential information we assigned a time span at which it is required, and an order of priority (results in Section 3.3). Information will be presented if and only if it is required at this particular point in time and is shown in descending order of priority.

This is achieved by creating a turn-by-turn-like interaction. Travelers continuously progress through the trip by steering from intermediate target to target and ultimately arrive at the destination. Only information required to reach the next intermediate target is conveyed to the user. The design of the GUI is intended to noticeably enhance users' public transport navigation experience in comparison with existing applications. It is supposed to aid the traveler's navigation and make processing routing information as easy as possible. Device-independent mockups presented in Figures 3(c) and 3(d) show an early stage of design.

3.1 Planning the Trip

A trip is the journey from a specified starting point (usually in the vicinity of the traveler's current location) to a destination point in a previously set time frame, using one or several different modes of travel. Although it is possible to add numerous other parameters to the plan, they are not included in the application for the sake of simplicity. Conceivable additional options for planning the trip, although increasing complexity, include specifying the desired date/time of arrival instead of departure, preference of direct connections or favored traveler profiles. Users of public transport navigation aids often are in need of public transit information regarding their current location, such as to quickly return home from a friend's house or when to take the train to get to university. Therefore incorporating the latest position update as a location parameter for the trip request is deemed appropriate.

Figure 3(a) shows the first depiction of the initial planning view to input the desired trip parameters. The time of departure is shown in a textual representation and changeable by using a mechanism or by input via a (virtual) keyboard. Finally, the modes of travel are to be chosen using a multiple-selection segmented control which lists all possibilities and in-



(a) Planing view. (b) Trip list view. (c) Trip view (bus). (d) Trip view (walking).

Figure 3: Early development sketches.

dicates their selection state. The bottom-most button labeled "Search" will then issue the request and fetch one or more trip plans. One reasonable improvement to the interface might be the use of meaningful icons instead of the textual description of the travel modes in the segmented control. This is due to the fact that reading a piece of plain text usually takes longer than recognizing a commonly used icon.

3.2 Selecting a Trip

After submitting the trip inquiry, the traveler has to be presented with the trips suggested. The mockup in Figure 3(b) depicts an early version of a table view whose rows each contain one of the provided itineraries. The times of beginning and arrival of the journey are mandatory attributes for people to judge the expediency of a proposed trip plan. Especially if there are multiple possibilities for the modal composition of a route, displaying the different transport modes during the selection progress distinguishes any unique features of the trip. Displaying icons unambiguously depicting the travel modes not only saves screen space but also makes it easier to comprehend the complete routing process at a glance. In addition the starting and destination locations which the traveler had previously specified are presented on top of the table.

3.3 Assisting the Trip

The trip view shown in Figures 3(c) and 3(d) represents the most significant component of the interface and is also where the traveler is likely to spend the highest amount of time while actually using the application. Being the main source of detailed information about the journey, the trip view needs to fulfill certain requirements. Most importantly, all required information needs to be quickly accessible and easily comprehensible. A traveler should be able to capture the information which is necessary for her navigation at one glance without any interaction. Following the principle of *Cascading Information* introduced in Section 1.2, it is assumed that the traveler does not require the complete trip information during her navigation process. Therefore, the whole trip plan can be subdivided into small entities containing only the features of a single step along the route. Using this distinction allows displaying these entities individually. A type of view is needed which enables comfortable progression through the route and is easy-to-use. Consequently, the interface will not feature a vertical scrolling view. Instead, it is intended all the data necessary for the corresponding stage of the trip are visible at once, eliminating any need to perform time-consuming repositioning of screen elements or going through all the information to find the respective piece of information.

Using Context Awareness

To present exactly the required information for the current situation, respectively the current step, knowledge about the current situation is required. The multitude of sensors included in current mobile devices allows for precise determination of user contexts and their surroundings, and therefore the situation of the traveler. By using the context data, e.g., time and date, the location and the type of transportation vehicle the application is enabled to ascertain whether a user has entered the vehicle she was headed for on her itinerary. Using this information we can automatically deduct which step of the planned route should be presented.

Since one step in the plan has a variety of attributes, which are potentially subject to change, the possibilities of how to visualize them are manifold. In general one can distinguish between purely textual

and graphical representations of data. As depicted in Figure 3(d), the mockup GUI contains a combination of both methods in order to take into account different user habits. Independent of the type of transport the general interface layout stays the same. The information is always displayed at the same area to enable the traveler to recognize relevant information at a glance.

Textual Elements

There are a few characteristics which are essential to the navigation procedure and therefore any major change to them is improbable. If the step required public transportation, the required characteristics, in descending order of priority, are (also shown in Figure 3(c)):

Departure and Arrival Stations. Naturally, the traveler needs to know the stations of the transit vehicle she intends to travel by. The description of the stop should not only include its name, but also some kind of information about which platform to enter. This holds especially true for large buildings like central stations or coach terminals.

Vehicle Information. Public transport vehicles are usually identified by a line number and direction headsign. Both are displayed to enable the traveler to find the vehicle easily.

Departure and Arrival Times. Knowing the respective departure and arrival times is a crucial factor for the traveler to catch the vehicle and leave it at the right time.

When providing on-foot navigation important characteristics include (shown in Figure 3(d)):

Direction. Walking navigation thrives on supplying the traveler with the direction she is supposed to walk in. While people usually tend to use relative directions like “left” or “slightly right” absolute directions based on points of the compass are a possible alternative.

Distance. Information about how long the current step will take is generally of great interest to a traveler. Since the individual walking speed may differ significantly between people, the better alternative is displaying the distance that has to be traveled.

Street Names. The traveler requires identifiers to exactly recognize the planned walking route she is supposed to follow. This is most easily accomplished by providing either the name of the street or an identifier of a nearby junction or a well-known public place.

Also, the final destination of the journey should always be visible so that the traveler is reassured she is traveling in the right direction. The top area provides sufficient space to display a text which slightly resembles the head sign of a bus or tram and therefore seem familiar to the traveler.

Graphical Elements

The graphical representation of trip data is primarily performed by a map view which displays the surroundings of the traveler and her route using an online map service as data source. All of the steps which are transmitted in the planned route are marked on the map and a path overlay approximately indicates the track that will be traversed during the journey. The traveler’s progress on the route is shown by highlighting the map marker which corresponds to the current step. A compass is shown on the map to add an additional easy to follow navigation cue. For the traveler to always be aware of her advancement we added a progress bar at the bottom of the display. To prevent distraction, all graphical elements should be kept simple and lean.

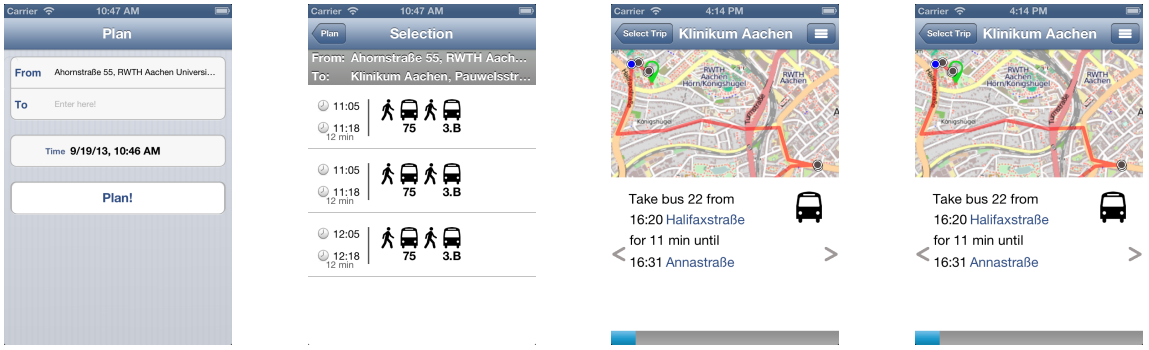
3.4 Back-end Requirements

Since the application is to be extensively evaluated in a user study, it would greatly benefit from real functionality instead of mockup-like dummy behavior. Therefore, the decision was made to employ a real back-end service capable of providing intermodal routing information for public transportation. Ideally, this service needs to include a variety of different travel modes as well as the possibility of extending the system. Most importantly the server is required to allow effortless import and use of own map and transit data in a standardized format so that user testing may be performed regionally with the appropriate information.

4 IMPLEMENTATION

This section describes the actual implementation based on our previously introduced approach. We begin with the process of drafting and assembling the prototype, including detailed explanations of the GUI followed by a brief description of the used back-end.

In the following we address the realization of the views of the prototype GUI. Since there was relatively little time between the creation of the mockups and the implementation of the application, and also because of the sophistication of modern mockup tools,



(a) Plan view. (b) Selection view. (c) Assistance view (bus). (d) Assistance view (walking).

Figure 4: Screenshots showing the prototype implementation.

the differences are not always apparent. Because of its popularity, stability, and the great amount of well-written documentation, Apple iOS 6 became the platform of choice for the development of the application prototype. In order to fully ensure compatibility with a real device, the application was also regularly tested on an iPhone 4.

4.1 The Planning View

The trip planning view presented in Figure 4(a) consists of *from* and *to* text fields, a date selector, and a *Plan!* button. Upon beginning text input a virtual keyboard is displayed alongside several possible text buttons. These buttons include a green position marker which resembles the traveler's position. By tapping it, the traveler starts a positioning process which uses the current geolocation, coordinates supplied by the device sensors to look up the current address using reserve geocoding. This trip planning may be used in two different scenarios: Either a traveler wants to make a request for an entirely new route, or she issued the command to reschedule an active trip. In the first case, the next step in the navigation chain is displaying the list of feasible itineraries provided by the backend server. If, on the other hand, a reschedule is supposed to happen, the application skips displaying the trip list and selects the first itinerary, optimistically assuming that it embodies the temporally closest possibility.

4.2 The Selection View

After receiving a response from the back-end server the trip information needs to be conveyed to travelers using the selection view in Figure 4(b). The shown table contains a variety of information; on the top of the list, the full description of both the origin as well as the destination location is presented. Each following row sums up the trip, containing on the left hand side the time of departure and arrival. Furthermore the full duration of the trip is written below. Next to the times the utilized modes of transportation are listed in an array of icons. Below the icons the line identifier (e.g., line number) is shown. To maintain consistency the icons from the planning view are reused throughout the application. By reducing the icon size after reaching a certain threshold an overflow is prevented.

4.3 The Assistance View

Presenting the stages of the trip to the traveler was one of the main challenges in designing and implementing the prototype GUI. Therefore drafting it included devising the functions which had to be employed in order to satisfy user needs while keeping a plain profile. The view grew in functionality and its features matured during the process of implementation.

Since the trip is split into several parts, a similar graphical representation is an obvious consequence. The trip plan is segmented into a number of steps which have to be subsequently completed in order to arrive at the destination. These steps are visualized on the map view by adding a position marker to the map layer at the geographical coordinates of the starting point of each step.

In addition to showing the route path and progress on the map, a *compass* was incorporated into the view. Using a rotatable image of a green arrow in the upper left corner, the traveler ought to always be aware of the direction in which her destination lies. This is achieved by computing the angle (i.e., the bearing) between the traveler's location and her destination. Utilizing iOS ability to not only monitor the position but also the heading, the application uses the bearing and the travelers heading to create and apply an affine transformation to the arrow image, thus rotating it. As a result, travelers are able to discern the correct direction of their destination at any given moment which is especially beneficial for walking direc-

tions. The concept of this orientation aid is derived from games, which often provide navigation instructions by displaying a compass. As the compass is only relevant for walking directions, it is disabled for other modes of transportation.

The part of the GUI displaying the textual representation presents either a step of the walking directions or a transit stage of the trip and contains a number of labels and one image view for depicting the current mode of the step, again reusing the mode icons from the planning view. If the step is part of the turn-by-turn walking directions it includes information about the absolute and relative direction, the street the traveler is supposed to follow, and the distance that will be traversed during this part of the journey (shown in Figure 4(d)). If, on the other hand, the view shows a transit stage, it contains details regarding the type of vehicle, the transit route, departure times and station as well as arrival time and name of the destination (Figure 4(c)).

At the bottom of the view, a progression indicator is placed which uses a bar style and is filled according to the progression in the trip. However, merely iterating over the steps does not accurately resemble the progression through the trip plan, because the route is generally not of a linear nature. Instead, the various parts of the journey each consume different amounts of time and therefore the advancement of the progress view was adjusted to precisely mirror the traveler's progress.

4.4 Back-end Realization

In order to generate some real functionality for the application prototype an existing routing solution was required. The optimal candidate in this scenario is an open source complete planning service to serve as a back-end which is capable of autonomously computing a route between two arbitrary locations. Also favorable are the possibilities of freely integrating a variety of maps sources as well as a fully compatible API to address the back-end server.

The OpenTripPlanner (OTP) project[2] is an open source platform which has developed a rather large community since its launch in 2009. While still in development, it already features intermodal trip planning with special support for bikes, allows the import of several internationally popular data types and offers a RESTful web API. OTP is mainly Java-based and consists of a server which performs routing and supplementary services, and a basic but effortlessly extensible web application enabling users to search for a route via their web browsers. Furthermore, the RESTful API is able to provide the same experience using either XML or JSON formatting, enabling access to the service with mobile clients.

The project employs the established A^* search algorithm for path finding when walking and the MOA^* algorithm for intermodal trips after the provided raw map data has been indexed by OTP's own map builder application. By default, all maps are obtained from the well-known OpenStreetmap (OSM) project, which distributes very up-to-date maps on an open source basis. To ensure compatibility, the developers use Google's General Transit Feed Specification (GTFS)[3] format for moving and importing transit data. The project includes a graph builder tool which is capable of analyzing any provided GTFS data feed and determining the required map excerpt, and automatically downloading the corresponding data from OSM. Afterwards, the program computes a large graph object containing all the correlated data ready to be used by OTP.

5 EVALUATION

For evaluating the prototype a user-oriented evaluation was applied.



Figure 5: Participant interacting with application prototype.

5.1 Methodology and Test Participants

The prototype application and the DB Navigator were evaluated in a lab user test (see Figure 5) in order to identify usability problems and have an insight in prospective users' assessment. The test encompassed a leading scenario and three tasks which had to be solved by interacting with the applications. The purpose of the leading scenario was to give the test tasks a contextual framing for supporting test participants in understanding the whole point of the test.

[2]http://opentripplanner.com

[3]https://developers.google.com/transit/gtfs

Scenario:

> You are a student in Aachen and you have a student ticket, which allows you to use all public transportation in your state. For this reason, you are a frequent user of Aachen's public transportation. You own an iPhone and use trip assistant apps to inform yourself about departure times.

After reading this, test participants were given the task instructions which focused to test the app's main functionality routing. To reduce the cognitive load and concentrate on the task processing, instructions for the next task were handed out after finishing the current one.

Task 1 was about routing from the actual location to a bus stop:

> It is 09:18 a.m. at the 3rd of August 2013. You are running through an unknown street. You have to go to the Audimax, a university building, to write an exam. The exam will start at 10:00 a.m. Use your app to find information about a bus ride to your exam. Use the integrated functionality to find your actual location.

Task 2 was about routing from a bus stop to an address:

> It is 03:03 p.m. at the 3rd of August 2013. Your performance during the exam was more or less successful. You are having a late lunch with your girlfriend Claudia and have to plan a trip from the Pontstrasse, Aachen to the home improvement store because you have to buy material for fixing a shelf for Claudia. The home improvement store is located in Roermonder Strasse 177, Aachen. You are again pressed for time because later you have an important appointment. Use your app to reach your destination by bus as quick as possible.

Task 3 was about routing from one address to another:

> It is 5:13 p.m. at the 3rd of August 2013. After you have carried out all tasks you have promised to Claudia you have to go quickly to your family for helping them to prepare a birthday party. At the moment, you are at Claudia's place in Oppenhoffallee 143, Aachen. Your parents live in Burtscheid, a suburb of Aachen, in the street Birkengrund 10. Aachen. You have to be there at 06:00 p.m. to set up a pavilion for the guests. You have to hurry to do this because the weather forecast has announced 88% rain probability. Use your app to reach your destination by bus as quick as possible.

The experimental design was varied for controlling position effects: Half of the test participants carried out the test tasks first with the prototype and then again with DB Navigator, the other half of test participants vice versa. During the test the test participants were instructed to apply the thinking aloud method which means to comment on their interaction with the apps. Verbal comments were recorded via audio and video. In addition, the interaction with the apps was recorded by a screen record software. The feedback of participants was collected with a questionnaire consisting of six sections:

Section 1 Demographic Data. Age, gender, profession, education.

Section 2 Mobility Profile. Possession of driver's license, use of means of transportation.

Section 3 Technology Use. Possession of electronic devices, frequency of use of devices, perceived ease of use of, electronic devices, experience with trip assistant apps.

Section 4 Assessment of the Prototype. Positive and negative comments, Additional questions about icons, size of map, visualization of progress.

Section 5 Final Rating. Application preference.

Ten test participants took part in the user test. Five were male, five female. Their age was $M = 23.9$ years, $SD = 3.03$ years. All of them were students. About their daily mobility profile they stated to go by foot (10), to ride the bus (5), to cycle (2), and to use the train (1). They were all using a smartphone (six iOS, one Android, one unknown, two others). Concerning mobility apps, nine used the DB Navigator, eight Google Maps, and three Navigon. Participants' experience with the DB Navigator was assessed as problematic because it impacts results of performance tests. However, this experience with several mobility apps could also be helpful for assessing the prototype's quality and utter specific optimization proposals.

5.2 Evaluation Results

The expenditure of time for solving tasks varied (see Table 1). According to the measured data, the processing of the tasks with the application prototype took longer than with DB Navigator. As mentioned before, nine of ten participants know and use DB Navigator and have learned to use the application to a certain extent. Therefore the result of the final rating for one app was predictable: seven participants preferred DB Navigator, two the prototype application, one neither of both apps.

Table 1: Mean of required time for solving tasks.

Application	Task 1	Task 2	Task 3
DB Navigator	1:13	2:44	4:22
Prototype	2:26	3:24	5:55

5.3 Participant Feedback

Participants gave very detailed feedback about the prototype visualization and features. They assessed positively that prototype was clearly and simply designed. For them it seemed to be easy to manage. No unnecessary gimmicks or icons were integrated. Furthermore, they appreciated the presentation of the map and the step-by-step support during the whole trip. The choice of means of transportation by marking icons was also a pro argument for the protoype application. Negative comments were predominantly related to missing features or information. For instance, participants missed the feature for defining the time of arrival/departure, choosing a bus stop from a selection, and a default setting of the territory, which should be considered (e.g. data for town, not for whole country). Moreover, information such as details of vehicles (e.g. bus number) was absent.

Technical problems were mentioned in connection with the GPS-based definition of the actual location. The range of participants' assessments concerning the size of the map was from "too small" to "ok". Also zooming was a little challenging for test participants. In general, the quality of icons rated as comprehensive. Two icons, bus and train, were named as too similar. The visualization of process progress within the app was also questioned. Three participants opted for page control dots, seven for a progress bar, two for displaying the percentage (multiple answers possible). Two participants stated that it was not clear, which kind of progress was visualized in the prototype distance, time, or steps within the step-by-step assistance.

5.4 Discussion

Based on the presented results our prototype application can be considered slightly inferior to the reference application DB Navigator. This outcome was expected; the majority of the test subjects had previous experience with the DB Navigator application. While this allows for potentially increased competence in operating the applications, it may have caused a biased outcome of the evaluation. The feedback highlighted the core usability issues of the interface, which severely influence the performance and satisfaction of users. While any problems identified by users during the tests are unfortunate, designing a flawless user interface right from the start is close to impossible. Last but not least, the employed lab test enviroment misses congnitive stress, which we believe will change the outcome.

6 CONCLUSIONS AND FUTURE WORK

This section discusses options of extending the prototype and the research topic in general and afterwards draws the conclusion of this paper.

6.1 Future Work

The presented mobile application is the foundation for a possibly holistic mobile application. However, the developed application is merely a prototype and therefore does not aspire to be complete or even a valid competitor on the market. During the process of implementation and beyond, many opportunities for extending and improving the design as well as the application itself were discovered or proposed by others. This section presents the most promising ideas and the circumstances of a possible realization.

Context Awareness

Although an approach to employ automatic context detection was described in this work, it was not included in the prototype. Instead, we opted for a manual selection of the current waypoint, respectively the current context as the focus of this work lies on the presentation of information. To reach the goal of a full-fledged mobile travel assistance solution, integrating an existing framework for automatic context is highly desirable. This would spare the user the need to manually select the next step of the journey by performing input on the device.

Indoor Navigation

A transit trip plan usually consists of many different stages with a multitude of varying transportation modes and changes of location. People often travel with one transit vehicle to a certain station and afterwards have to transfer to another vehicle. This happens especially often at stops in large facilities like central train stations or airports. Depending on the size of these buildings, the number of different transit modes they support travelers may be overwhelmed by their complexity. Users who visit the station in question for the first time are particular at risk of going

astray in the building complex and therefore missing their connection. As GPS positioning is only available outdoors, the traditional means of locating the user fail to operate properly. In addition to this, there usually is no map data available for indoor navigation. Therefore we plan the incorporation of our previous work supporting pedestrian indoor navigation into the prototype.

Social Networking

Social networking is an aspect which increasingly finds its way into many everyday activities and often manages to make ordinary tasks more appealing. By adding social components like interaction with friends by e.g., sharing, commenting or recommending itineraries, the application might enhance the user experience and consequently create a new incentive for repeatedly using it.

Real-time Transit Data

Because the application is supposed to aid the traveler during her trip and guide her through problematic situations, using and providing real-time information about transit schedules is a crucial extension. By including this data in the application it would become possible to inform travelers of any unforeseen circumstances like delays or even cancellations. Using the application, the traveler could then react to the delay and try to search for an alternative way.

Field Test

After discerning and correcting a large amount of the design errors during the first user test in the laboratory, the logical consequence is a new iteration of usability testing. The initial user study was mainly intended to be a first step towards adjusting the application to user needs. However, the testing environment was unnatural considering the performed tasks, because people normally use public transit navigation either on the road or before starting a journey.

In order to further evaluate the application prototype more realistically a field test will be performed, which requires users to actually embark on a real trip. During this pre-planned journey the users and their reactions will be closely monitored in order to determine the amount of cognitive capacities that is actually consumed by using the application for orientation in a realistic scenario.

6.2 Conclusion

This paper presented the attempt to design a novel Graphical User Interface for planning and routing through multimodal transit trips. The interface prototype was designed to alleviate the issues of travelers and guide them while they embark on their journey. The first draft of the GUI was produced employing the principle of cascading information and classic user interface design methods, determining the scope of this work.

Based on the derived concepts and drafts, a prototype mobile application was developed for Apple iOS. The trip planning was based on a backend server running OpenTripPlanner. The developed application was finally evaluated in a laboratory test user study in order to obtain information about usability, design issues and the comparability of the prototype with established applications for public transport navigation. The results of this study revealed the crucial usability problems and provided detailed feedback about the aspects of the GUI. While the prototype lacks the developmental stage of a commercial application and offers room for improvement, the positive user feedback showed that presented approach is reasonable and worth further attention.

ACKNOWLEDGEMENTS

The authors would like to thank Teresa Schmidt and Christian Paul for conducting the user study and their former colleague Paul Heiniz for productive discussions and his work on indoor navigation. The authors would also like to thank ASEAG AG and IVU Traffic Technologies AG for supplying the required timetable data.
This work was supported by the German Federal Ministry of Economics and Technology[4]:
(Grant 01ME12052 econnect Germany).

REFERENCES

Baus, J., Krüger, A., and Wahlster, W. (2002). A Resource-Adaptive Mobile Navigation System. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI 2002)*, pages 15–22, New York, NY, USA. ACM.

Biagioni, J., Agresta, A., Gerlich, T., and Eriksson, J. (2009). TransitGenie: a Context-aware, Real-time Transit Navigator. In *Proceedings of the 7th ACM*

[4]Bundesministerium für Wirtschaft und Technologie (BMWi) http://www.bmwi.de/

Conference on Embedded Networked Sensor Systems (SenSys 2009), pages 329–330, New York, NY, USA. ACM.

Biagioni, J., Gerlich, T., Merrifield, T., and Eriksson, J. (2011). EasyTracker: Automatic Transit Tracking, Mapping, and Arrival Time Prediction using Smartphones. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys 2011)*, pages 68–81, New York, NY, USA. ACM.

Chowaw-Liebman, O., Christoph, U., Krempels, K.-H., and Terwelp, C. (2010). Indoor Navigation Approach Based on Approximate Positions. In *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN 2010)*, pages 15–17.

Christoph, U., Krempels, K.-H., von Stlpnagel, J., and Terwelp, C. (2010). Automatic Context Detection of a Mobile User. In *Proceedings of the International Conference on Wireless Information Networks and Systems (WINSYS 2010)*, pages 1–6.

Corneliussen, H. and Rettberg, J. (2008). *Digital culture, play, and identity: A World of Warcraft reader*. MIT Press.

Deterding, S., Sicart, M., and Nacke, L. (2011). Gamification. using game-design elements in non-gaming contexts. *Extended Abstracts on Human Factors in Computing Systems*, pages 4–7.

Heiniz, P., Krempels, K.-H., Terwelp, C., and Wüller, S. (2012). Landmark-based Navigation in Complex Buildings. In *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN 2012)*, pages 1–9.

Keller, C., Korzetz, M., Khn, R., and Schlegel, T. (2011). Nutzerorientierte Visualisierung von Fahrplaninformationen auf mobilen Gerten im ffentlichen Verkehr [User-oriented Visualization of Train Schedule Information on Mobile Devices for Public Transportation]. In *Mensch & Computer 2011*, pages 59–68. Oldenbourg Wissenschaftsverlag GmbH.

May, A. J., Ross, T., Bayer, S. H., and Tarkiainen, M. J. (2003). Pedestrian navigation aids: information requirements and design implications. *Personal and Ubiquitous Computing*, 7(6):331–338.

Pessemier, T., Dooms, S., and Martens, L. (2013). Context-aware recommendations through context and activity recognition in a mobile environment. *Multimedia Tools and Applications*, pages 1–24.

Rehrl, K., Leitinger, S., Bruntsch, S., and Mentz, H.-J. (2005). Assisting orientation and guidance for multimodal travelers in situations of modal change. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportations Systems (ITSC 2005)*, pages 407–412.

Statista GmbH (2012). App Monitor Deutschland November 2012. Technical report, Hamburg.

Vieira, V., Caldas, L. R., and Salgado, A. C. (2011). Towards an Ubiquitous and Context Sensitive Public Transportation System. In *Proceedings of the 4th International Conference on Ubi-Media Computing (U-Media 2011)*, pages 174–179.

Yu, H. and Lu, F. (2012). Advanced multi-modal routing approach for pedestrians. In *Proceedings of the 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet 2012)*, pages 2349–2352.

Zheng, Y. and Xie, X. (2011). Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology*, 2(1):2:1—-2:29.

Evaluating Impact of Cross-platform Frameworks in Energy Consumption of Mobile Applications

Matteo Ciman and Ombretta Gaggi
Department of Mathematics, University of Padua, Via Trieste 63, 35121 Padua, Italy
{*mciman, gaggi*}*@math.unipd.it*

Keywords: Mobile Applications Development, Cross-platform Frameworks, Energy Consumption.

Abstract: In this paper we analyze energy consumption of mobile applications using different smartphones sensors, e.g., GPS, accelerometer, etc., and features, e.g., acquiring video or audio from the environment. In particular, we have studied how the use of frameworks for mobile cross-platform development may influence the amount of required energy for the same operation. We use an hardware and software tool to measure energy consumption of the same application, using different sensors, when developed natively or using two frameworks, Titanium and PhoneGap. Our experiments have shown that frameworks have a significant impact on energy consumption which greatly increases compared to an equal native application. Moreover, the amount of consumed energy is not the same for all frameworks.

1 INTRODUCTION

With the advance of mobile technologies, modern mobile devices are equipped with an ample set of sensors, like GPS, accelerometer, light sensor, etc. The presence of these sensors allows the implementation of more and more attractive applications, which can analyze the surrounding environment to better answer to the user's need. As an example, applications can use GPS to provide information about interesting places near the user, the accelerometer to understand the current user activity (e. g., if he/she is walking, riding a bike or using a car), and the light sensor to adapt the screen brightness.

Acquiring data from all these sensors have a cost in terms of energy. The software implementing the acquisition of data from sensors and the adaptation process must conserve battery power since energy is a vital resource for mobile computing and smartphones may operate for days without being recharged.

Battery life is, in fact, a critical performance and user experience metric on mobile devices. As stated by Bloom et al. (Bloom et al., 2004), battery life is one of the most important aspects considered by users when dealing with mobile devices, and most of the time users request a longer battery life. For this reason, energy consumption is an extremely important factor to consider when developing applications for smartphones. Unfortunately, it is difficult for developers to measure the energy used by their apps both before and after its implementation.

In this paper, we analyze energy consumption of mobile applications which acquire data using different smartphones sensors, e.g., GPS, accelerometer, etc., and features, e.g., acquiring video or audio from the environment. We compare the request in terms of power consumption of different sensors, with different samples frequency. We use the Monsoon Power Monitor (Monsoon Solutions Inc., 2013) during our experiments to reduce the impact of external events. We must note here that this tool requires to have direct access to the battery, therefore cannot be used with iOS devices without opening them.

Moreover, we have studied how the use of frameworks for cross-platform development may influence the amount of required energy for the same operations. Our experiments have shown that frameworks have a significant impact on the total amount of consumed energy, since an application developed using a cross-platform framework for mobile development can consume up to 60% more energy than an equal native application, which means, a strong reduction in terms of battery life. Moreover, the amount of consumed energy is not the same for all frameworks. This means that power consumption can be one of the factors to be considered in the choice between native implementation and using a framework, or in the choice between two different frameworks.

The paper is organized as follows: related works are discussed in Section 2. Section 3 describes the

frameworks for cross-platform mobile development already present in literature. The experiments made are presented in Section 4. We conclude in Section 5.

2 RELATED WORKS

Other works in literature analyze power consumption of mobile applications but they usually do not cover all sensors/features available on smartphones, and do not consider the use of cross-platform framework for development of mobile applications.

(Balasubramanian et al., 2009) measure energy consumption of mobile networking technologies, in particular 3G, GSM and WiFi. They find out that 3G and GSM incur in tail energy overhead since they remain in high power states also when the transfer is complete. They developed a model for energy consumed by networking activity and an efficient protocol that reduces energy consumption of common mobile applications.

(Thompson et al., 2011) propose a model-driven methodology to emulate the power consumption of smartphone application architectures. They develop SPOT, *System Power Optimization Tool*, a tool that automates code generation for power consumption emulation and simplifies analysis. The tool is very useful since it allows to estimate energy consumption of potential mobile architecture, therefore *before* its implementation. This is very important since changes after the development can be very expensive. Moreover the tool is able to identify which hardware components draw significantly more power than others (e.g, GPS).

(Mittal et al., 2012) propose an energy emulation tool that allows to estimate the energy use for mobile applications without deploying the application on a smartphone. The tool considers the network characteristics and the processing speed. They define a power model describing different hardware components and evaluate the tool through comparison with real device energy measurements.

PowerScope (Flinn and Satyanarayanan, 1999a; Flinn and Satyanarayanan, 1999b) is a tool to measure energy consumption of mobile applications. The tool calculates energy consumption for each programming structure. The approach combines hardware instrumentation to measure current level with software to calculate statistical sampling of system activities. The authors show how applications can modify their behavior to preserve energy: when energy is plenty, the application allows a good user experience, otherwise it is biased toward energy conservation.

AppScope (Yoon et al., 2012) is an Android-based energy metering system which estimates, in real-time, the usage of hardware components at a microscopic level. AppScope is implemented as a kernel module and provides an high accuracy, generating a low overhead. For this reason, the authors also define a power model and measure energy consumption with external tools to estimate the introduced error, which is, in the worst case of about 5.9%.

Eprof (Pathak et al., 2012a), is a fine-grained energy profiler for mobile apps, which accurately captures complicated power behavior of smartphone components in a system-call-driven Finite State Machine (FSM). *Eprof* tries to map the power drawn and energy consumption back to program entities. The authors analyzed the energy consumption of 21 apps from Android Market including AngryBirds, Android Browser, and Facebook, and they found that third party advertisement modules in free apps could consume up to 65-75% of the total app energy, and tracking user data (e.g., location, phone stats) consumes up to 20-30% of the total energy. Moreover, smartphone apps spend a major portion of energy in I/O components such as 3G, WiFi, and GPS.

Pathak et al. (Pathak et al., 2012b) study the problem of no-sleep energy bugs, i. e., errors in energy management resulting in the smartphone components staying on for an unnecessarily long period of time. They develop a static analysis tool to detect, at compile-time no-sleep bug in Android apps.

In other papers, the authors compare different framework for cross-platform mobile development according to a set of features. (Heitkötter et al., 2013) compare jQuery Mobile (Firtman, 2012), Sencha Touch (Sencha Inc., 2013), The-M-Project (Panacoda GmbH., 2013) and Google Web Toolkit combined with mgwt (Kurka, 2013) according to a particular set of criteria, which includes license and costs, documentation and support, learning success, user interface elements, etc. They conclude that jQuery Mobile is a good solution for simple applications or as first attempt in developing mobile apps, while Sencha Touch is suited for more complex applications.

(Palmieri et al., 2012) evaluate Rhodes (Motorola Solutions, Inc, 2013), PhoneGap (Apache Software Foundation, 2013), dragonRAD (Seregon Solutions Inc., 2013) and MoSync (MoSync Inc., 2013) with particular attention to the programming environment and the APIs they provide. The authors provide an analysis of the architecture of each framework and they conclude highlighting Rhodes over other frameworks, since this is the only one which supports both MVC framework and web-based services.

(Ciman et al., 2014) evaluate different cross-platform frameworks, i.e. Phonegap, Titanium,

jQuery Mobile and MoSync, with the focus on applications with animations, i.e. games. Besides the standard evaluation parameters like IDE, debug tools, programming complexity, etc., they evaluate more mobile and games-related aspects like APIs for animations, mobile devices supported, support for Native User Interface, performances etc. They conclude that, according to the actual state of art of the frameworks, Titanium is the best framework, since it supports animations and transitions, and its performances are good even in case of complex applications.

Issues about performances are discussed in (Charland and Leroux, 2011). They stated frameworks based on web technologies experience a decrease in performances when the applications implement transition effects, effects during scrolling, animations, etc., but this problem affects essentially games, while the loss in performances are unnoticeable in business application, i. e., applications which support business tasks.

All the addressed works make a critical analysis of the chosen frameworks according to criteria which never include power consumption. In some case they include performances which are considered in terms of user experience. In this paper we want to study how the use of a cross-platform framework for mobile device may affect the energy consumption of the final application with respect to native development.

3 FRAMEWORKS FOR CROSS-PLATFORM MOBILE DEVELOPMENT

According to Raj and Tolety (Raj and Tolety, 2012), frameworks for cross-platform development can be divided into four approaches: *web*, *hybrid*, *interpreted* and *cross compiled*. They highlight strength and weakness of each approach, concluding that a preferred solution for each kind of application does not exist, but the decision about which framework to use should be made considering the features of the application to be developed. To help the developers in this decision, they provide a matrix which shows which are the best choices to develop a specific feature. This classification can help to correctly interpret the results of our tests.

The *Web Approach (WA)* allows the programmer to develop a web application using HTML, CSS and Javascript. Given the new emerging features of HTML5 and CSS3, these technologies allow the creation of rich and complex applications, therefore their use cannot be considered a limitation. The application can be accessed through a mobile device using only its integrated browser, connecting to the right public internet address. Figure 1 shows this approach. In this way, a full devices support is guaranteed (limited only by the maturity of the native browser respect to the standard technologies specifications), but the problems arise when the application needs to access to the smartphone's sensors, because this feature is not provided for web applications. Moreover, the user does not interact with a mobile application, but he/she must run the browser to access a web page containing the web application. An example of framework using the WA is *jQuery Mobile* (Firtman, 2013).

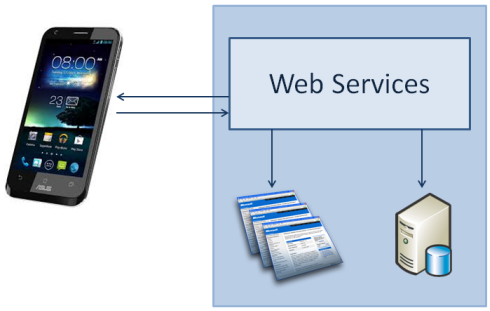


Figure 1: Architecture of an application using the WA.

The *Hybrid Approach (HA)* is a middle way between *native* and *web approach*. In this case, the application operates in two different ways. It uses the webkit rendering engine to display controls, buttons and animations. The webkit engine is therefore responsible to draw and manage user interface objects. On the other site, the framework and its APIs provide access to device features and use them to increase the user experience. In this case, the application will be distributed and installed and appears on the user device as native mobile applications, but the performances are often lower than native applications because its execution requires to run the browser rendering engine. An example of this kind of frameworks is PhoneGap, also known as Apache Cordova (Apache Software Foundation, 2013), and the architecture of the final application is shown in Figure 2.

With the *Interpreted Approach (InA)*, the developer has the possibility to write the code of the application using a language which is different from languages natively supported by the different platforms. An example can be Javascript: developers who already knows this language are simply required to learn how to use the APIs provided by the framework. When the application is installed on the device, the final code contains also a dedicated interpreter which is used to execute the non-native code. Even in this case, the developer has access to the device features (how

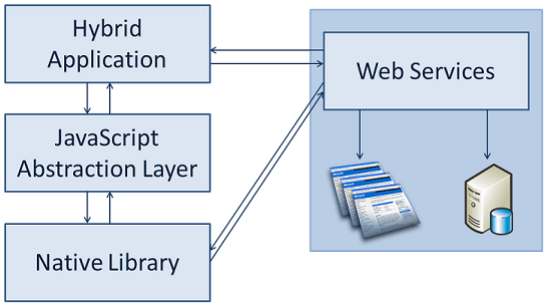


Figure 2: *Hybrid approach* application architecture.

many and which features depend on the chosen framework) through an abstraction layer. This approach allows the developer to design a final user interface that is identical to the native user interface without any additional line of code. This approach can reach an high level of reusable code, but can reduce a little the performances due to the interpretation step. Titanium (Appcelerator Inc., 2013a) is an example of this kind of framework and Figure 3 shows the architecture of the framework.

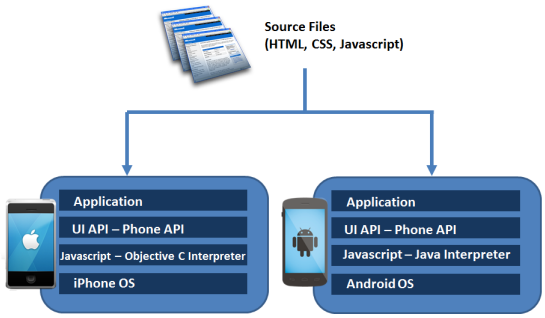


Figure 3: *Interpreted Approach* architecture.

The last approach, the *Cross Compiled Approach (CCA)*, lets the developer write only one application using a common programming language, e. g. C#. Frameworks which follow this approach generate, after compilation, different native applications for different mobile platforms. The final application uses native language, therefore can be considered a native mobile application to all intents and purposes. Therefore, this approach lets the programmer have full access to all the features provided by smartphones and the performances can be considered similar (even if not equal) to the native approach for simple application. In fact some tests have shown that for complex applications the native solution is better since the generated code gives worst performances of the resulting application, compared to code written by a developer. An example of this framework is Mono (Monologue Inc., 2013).

4 EXPERIMENTS

Our goal is to provide objective information about energy consumption of the most common sensors with which smartphones are usually equipped. To perform the best measure of energy consumption, it is important to avoid influences from external factors. A simple possibility is to run a background application that measures battery power at fixed intervals of time. This solution is adopted by some of the related works discussed in Section 2, but clearly introduces an overhead and must consider the possibility that other external events like call, messages, network problems and connectivity may influence energy consumption. Moreover these external events are almost unpredictable and unlikely reproducible, thus leading to unequal test sets.

4.1 Study Setup

For the reason mentioned before, during our experiments we use the Monsoon Power Monitor (Monsoon Solutions Inc., 2013). The hardware device comes with a software tool, the Power Tool, which gives the possibility to analyze data of energy consumption of any device that uses a single lithium battery. The information retrieved are energy consumption, average power and current, expected battery life, etc. These data are extremely important because can help the developer to understand which tasks use most of the energy of the battery, for how much time, etc. Moreover, it helps to analyze and improve the application code according to its power consumption. Using data acquired by Power Monitor, it is possible to understand if and where some energy can be saved increasing battery life, an extremely important aspect when working with mobile devices. Figure 4 shows the hardware setup of the system.

Our goal is to compare energy consumption of applications developed in a native way or using a

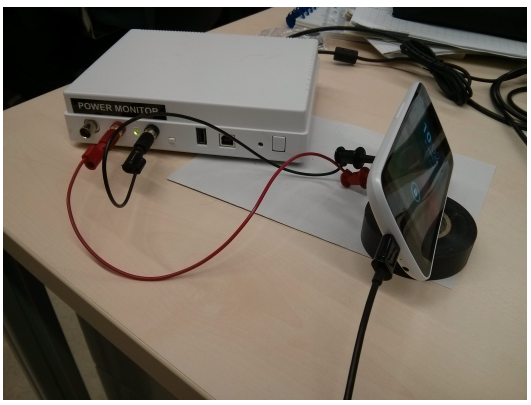


Figure 4: Hardware setup of the system.

cross-platform framework, and to compare performances, in terms of energy consumption, among different frameworks. The application used for our analysis allows to choose between several sensors and the frequency with which they retrieve data and update the user interface showing the retrieved data. This application can be considered a sort of lower bound for applications behavior in terms of energy consumption, because usually applications that collect data from sensors, perform several computation before updating the user interface.

As explained before, to use the Power Monitor tool it is necessary to have direct access to the battery of the smartphone, in order to be able to connect the entire system and to measure the consumed energy. For this reason, we made our comparison using an Android device with removable battery, because the iOS devices do not allow to have direct access to the battery in a safe way.

To compare energy consumption of mobile cross-platform frameworks, we select Titanium and PhoneGap. As already explained in Section 3, these two frameworks belong to two different categories. Titanium belongs to the *interpreted* frameworks, while PhoneGap follows an *hybrid* approach. In this way, first of all we can compare the native approach and its consumption with two different cross-platform frameworks. Then, we can even compare two different frameworks, getting information about the less expensive, in terms of power consumption, information that can influence the framework choice when considering mobile application development.

Our test smartphone is a Samsung Galaxy i9250. The (theoretical) capacity of the battery is 1750mAh. Even if this smartphone is not an up-to-date device, what is important for us is if the usage of a cross-platform framework has negative effects on energy consumption, and in which measure, for applications that use smartphone's sensors. With this question in mind, is easy to see that what we want to know, after our experiments, is how much (in percentage) the energy consumption increases. Even if characteristics and performances of sensors may vary between different smartphones, our tests are performed on the same device and comparisons are made on the percentage increase of energy consumption, to limit the influence of the chosen device on our final results. We must note here, that many devices, although different in terms of brand and model, share the same hardware components for sensors.

To make our tests, we used three different applications: a native mobile application, one built with PhoneGap and another one used to test the Titanium framework. We developed from scratch the application used to test the native solution and the Phonegap framework. These applications let the user chooses the sensor to test and the sampling frequency. Instead, we tested the Titanium framework using the Kitchen Sink app provided by Appcelerator (Appcelerator Inc., 2013b), a sample application built to show the different APIs and features provided by the framework.

All the applications retrieve data from the accelerometer, the compass and the GPS tests and display this information simply as a numerical value with a label. To test the microphone API, we record audio from the microphone and save it on the device as a 3GPP file. To test energy consumption due to the use for the camera, all the developed applications show the images captured from the camera for a predefined time interval, and take a picture at its end. Each test lasted two minutes. Previous experiments show that, after an initial interval, the duration of a test does not influence the final value of the expected battery life, and we chose this amount of time to get a stable final value from the system. The version of the Android API was 4.2.2, PhoneGap was version 3.0.0 while Titanium SDK was version 3.2.

4.2 Results

In this section we report the final results of our analysis. We analyze the most common sensors which can be used through the APIs provided by at least one the two analyzed frameworks. For each sensor, our application (either native or developed using a cross-platform framework) starts to acquire data coming from the selected sensor and display them on the user screen.

We made our experiments keeping the screen on, since this situation is much more adherent to reality where it's very difficult that an user interact with an application keeping the screen off: consider, as an example, Google Maps, which uses the GPS sensor to capture the position of the user and to give the correct directions.

Measures are repeated several times, in order to get a mean final value of each test. The main value we are interested in is the consumed energy, which shows how much energy is used by that particular task during this two minutes test. The smartphone was in flight mode (thus unable to receive any phone call or message that could arbitrarily increase energy consumption) and with WiFi and Bluetooth connectivity are turned off.

To get an idea about energy consumption of an application, we need a base value to compare with. For our purpose, we decided to measure the consumed

energy when the smartphone is turned on, with the screen on and without any running applications. This is the base value for our analysis. Clearly, it is quite impossible to measure the same value from another smartphone, but, since our discussion does not deal with absolute values of energy consumption, but with the increment in energy consumption in term of percentage, this initial value is important for our computation. Using the test smartphone, the consumed energy in two minutes is 6047,31 μAh. Moreover, the analysis of the increment in energy consumption reported in percentage helps to give results which are not tightly related to the chosen hardware.

Table 1 shows the results of our experiments, i. e., it compares how the consumed energy increases when the smartphone uses its sensors with a native application, or with applications developed using the two frameworks, PhoneGap and Titanium. We must note here that native development provides full access to all the available sensors of the device, while the number and type of supported sensors may vary between different cross-platform frameworks, depending on the state of the art of the frameworks themselves. For example, Titanium provides access to the microphone and record audio data only for iOS devices and not for Android devices, while PhoneGap supports this features for iOS, Android and Windows Phone.

The columns Δ of Table 1 show how the consumed energy increases compared to the consumed energy of our base value, i.e. the smartphone with the screen on, without running applications. This value gives an idea of how much more expensive is to perform a particular task among our initial idle state.

As it is easy to see, the differences between power consumed by the native app and the solutions developed with a framework for cross-platform development is very high.

Let us begin the analysis with the comparison between applications which do not capture any data from sensors. The purpose of this test is essentially to investigate if the adoption of a cross-platform framework instead of a native development requires more energy consumption simply to "show" the application without any computation behind. The results are shown in the first row of Table 1 denoted with the label "Only App". As we can see, the energy consumption increases of about 7% for PhoneGap and slight more than 2% for Titanium, i. e., the adoption of a cross-platform framework produces, basically, a little more expensive applications in terms of power consumption, and this increment is not equal for all frameworks.

Considering applications that use smartphones sensors to retrieve data, we can measure differences only for sensors that are supported at least from one of the two frameworks. The measurements made show that the usage of the accelerometer requires about 60% more energy using the PhoneGap application instead of the native one. This is an extremely high value, which means that the usage of this sensor in a cross-platform application is really a battery consuming task that can decrease user experience. Therefore, if the application needs to retrieve data from the accelerometer it is necessary to consider if it would be better to develop different, more performing, native applications for each platform.

The result reported for the Titanium framework needs a particular attention. Data reported on the table could lead to the wrong conclusion that it cost less energy to retrieve data with the Titanium framework in respect to the PhoneGap solution. The reason for this lower consumption is that the update frequency for the native (and the PhoneGap) application is about a sample every 64ms, while the Titanium frequency is about a sample every 600ms. To read data and update the user interface with a lower frequency (about 10 times lower) clearly reduces the amount of required energy. To be able to compare the performances of the Titanium framework, we used the PhoneGap application to acquire data at the right frequency (600ms) and we compared the final results (see Table 2). What we got is that, with the same update frequency, the delta of PhoneGap is about +40%, while for Titanium is about +98%. This means that, if compared together, the PhoneGap framework works better than the Titanium framework (about 60%).

All the cross-platform frameworks perform worst than the native solution also for data acquisition from compass and GPS: about 45% more power for acquisition from compass and about 10% more power for acquisition from the GPS using the PhoneGap framework; a little less than 5% more power for acquisition of data from GPS using Titanium, compass is not supported by Titanium yet.

The only sensor for which the difference in power consumption is very low is the use of camera to take pictures. In this case, the amount for energy consumed with a native application and with an application developed with a cross-platform framework differs of 8% - 12%, which is very low if we consider the total amount of energy required to use the camera (see the "Camera" row of Table 1). This behavior can be explained because in this case the cross-platform application makes only one call to the system API, and waits from the system the result (an image) and so the difference between the native and the cross-platform solution is really low.

Table 1: Energy consumption comparison between native applications and apps developed with a framework for cross-platform development. Note that data marked with (*) use a different updating frequency.

	Native		PhoneGap		Titanium	
Sensor	**Consumed Energy (μAh)**	**Δ (%)**	**Consumed Energy (μAh)**	**Δ (%)**	**Consumed Energy (μAh)**	**Δ (%)**
Only App	7705,54	+27,42%	8130,85	+34,45%	7860,97	+29,99%
Accelerometer	9179,99	+51,80%	12849,82	+112,49%	11972,16*	+97,97%*
Compass	9489,85	+56,93%	12124,6	+100,50%	-	-
Microphone (Rec)	8120,92	+34,29%	8404,71	+38,98%	-	-
GPS	9301,48	+53,81%	9947,60	+64,50%	9577,27	+58,37%
Camera	21857,38	+261,44%	22347,52	+269,54%	22576,45	+273,33%

A similar situation takes place for data acquired from the microphone (although test data are not available for the Titanium framework) and from the GPS. In the last case the difference, in terms of energy consumption, ranges between 5 and 11%.

Comparing the frameworks PhoneGap and Titanium, the overhead of energy consumed introduced by the two frameworks is particularly different when the user interface has to be updated frequently. This difference comes from the nature of the two frameworks. As already mentioned in Section 3, PhoneGap belongs to the *hybrid* family, while Titanium to the *interpreted* frameworks. This means that if we compare the two different platforms in terms of performances and energy consumption, the *hybrid* application is better, since the consumed energy by this application is lower, meaning that the overhead introduced is lower.

Another possibility to compare different development frameworks, and in particular how sensors usage affects application performances and energy consumption, would be to test these applications without updating the values of the retrieved information on the screen or turning off the screen while performing operations. In this case, the differences between the native solution and the cross-platform solutions when acquiring data from accelerometer or compass are much more lower, about 20%. Unfortunately, this is only a theoretical result. In fact, every application that retrieves data from sensors does not use this raw data immediately, but it usually elaborates the data and updates the user interface accordingly. Therefore, to use this theoretical results to promote the use of cross-platform framework would not adhere to the reality because we would not consider two extremely important parts of the applications, i. e., data elaboration and user interface management.

If we compare together the same cross-platform application, what we can note is that the difference in terms of consumed energy to show or not to show data from sensors is about 80% for both the framework. This essentially means that, without concern on the chosen cross-platform framework, the most expensive task for it is to update the User Interface.

Unlike Titanium, PhoneGap allows the developer to decide the frequency to retrieve data from sensors. This is an extremely important option, because lets the developer to define the right update frequency depending on the target application and the needs of data. This possibility is available for both the accelerometer and the compass sensor. In this cases, we made several test at predefined update frequencies (60ms, 150ms, 300ms, 600ms) to see how, and in which measure, changing the update frequency affects energy consumption. The results are shown in Table 2. As it is easy to see, if the update frequency decreases, the consumed energy decreases, with a difference that can reach 70% between 60ms and 600ms update frequency. This means that the developer has to pay extremely attention when developing a cross-platform application, because if the user experience do not requires an extremely fast update, it is useful to reduce the update frequency of sensor data in order to save energy, and so battery life.

5 CONCLUSION

Due to the diffusion of different smartphones and tablet devices from different vendors, the cross-platform development approach, i. e., to develop only one single application and distribute it to different devices and mobile platforms, is growing and becoming extremely important. This cross-platform development incorporates several approaches, i. e., *web*, *hybrid*, *interpreted* or *cross compiled*. All these approaches have several positive and negative aspects, either from a developer or a user point of view.

Many papers address the problem of how to choose the best framework for the development of a particular application, but, to the best of our knowledge, no one considers the power consumption as one of the key issue to make a correct choice.

In this paper we analyze the influence of the different approaches on energy consumption of mobile

Table 2: Consumed energy using different sampling frequencies to capture data with PhoneGap.

	Consumed Energy increase (%)			
Sensor	**60ms**	**150ms**	**300ms**	**600ms**
Accelerometer	+112,49%	+70,06%	+49,84%	+40,25%
Compass	+100,50%	+75,31%	+52,92%	+46,62%

devices, in particular for the *Hybrid* (PhoneGap) and the *Interpreted* (Titanium) approach. We compared the performances of a simple application, built with a particular framework, that retrieves data from sensors and updates the user interface, to a native application with the same behavior, to measure differences in terms of power consumption. Despite other previous analysis, we do not use a software to measure energy consumption since it introduces a not valuable overhead; for this reason we used the Monsoon Power Tool which allows to measure, through hardware links, consumed energy and to understand how this consumption increases with the two frameworks.

Our comparison shows that, visualization of data (business applications) perform better on the *interpreted* approach (Titanium), that can be a good solution for simple applications that do not retrieve data from sensors, e. g., on-line shopping, home banking, games which do not use accelerometer, etc. Moreover, the two analyzed frameworks, PhoneGap and Titanium, have very similar performances for Camera and GPS.

A particular sensor needs specific discussion. Even if Titanium seems to be the right choice if energy consumption is a key issue, our experiments have shown that it fails in case of retrieval of data from accelerometer for two reasons: the available API does not allow to impose an update frequency, and compared to other framework with the same update frequency it consumes about 60% more energy than PhoneGap. This result derives from the *Interpreted approach*: the necessary interpretation step can be extremely expensive and lead to more energy consumption. We must note here that this sensor consume a lot of energy, therefore if the application make a strong use of the accelerometer, the developer has to consider also the native solution.

The results that we got show that a cross-platform approach involves an increase in terms of energy consumption that is extremely high, in particular in presence of an high usage of sensors data and user interface updates. This increase can vary even in order of about 60%, meaning that it is important to choose the right framework to preserve the battery duration and to avoid negatively affecting user experience with usage of a cross-platform framework during development. In fact, several research studies have shown that energy consumption and battery life are the most important aspects considered by mobile devices users.

The results provided are clearly related to the state of art of the framework under examination at the time of writing. This means that, they can change in the future according to the improvements of the frameworks.

As future works, we plan to increase the total number of analyzed frameworks and to add HTML5 applications in our comparisons, trying to cover all the different approaches. Moreover, we will try to follow the development of the different frameworks in order to reach a complete analysis of the different sensors provided by the smartphones and their consumption in user applications.

REFERENCES

Apache Software Foundation (2013). Phonegap, http://phonegap.com/.

Appcelerator Inc. (2013a). Titanium, http://www.appcelerator.com/platform/titanium-platform/.

Appcelerator Inc. (2013b). Titanium Mobile Kitchen Sink Demo, https://github.com/appcelerator/KitchenSink.

Balasubramanian, N., Balasubramanian, A., and Venkataramani, A. (2009). Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet Measurement Conference*, IMC '09, pages 280–293.

Bloom, L., Eardley, R., Geelhoed, E., Manahan, M., and Ranganathan, P. (2004). Investigating the relationship between battery life and user acceptance of dynamic, energy-aware interfaces on handhelds. In *Proceedings of the International Conference Human Computer Interaction with Mobile Devices & Services*, pages 13–24.

Charland, A. and Leroux, B. (2011). Mobile application development: web vs. native. *Communications of ACM*, 54(5):49–53.

Ciman, M., Gaggi, O., and Gonzo, N. (2014). Cross-platform mobile development: A study on apps with animations. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC'14.

Firtman, M. (2012). *jQuery Mobile: Up and Running - Using HTML5 to Design Web Apps for Tablets and Smartphones*. O'Reilly Media.

Firtman, M. (2013). jquery mobile, http://jquerymobile.com/.

Flinn, J. and Satyanarayanan, M. (1999a). Energy-aware adaptation for mobile applications. In *Proceedings of*

the seventeenth ACM Symposium on Operating Systems Principles, SOSP '99, pages 48–63.

Flinn, J. and Satyanarayanan, M. (1999b). Powerscope: A tool for profiling the energy usage of mobile applications. In *Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications*, WMCSA '99, Washington, DC, USA. IEEE Computer Society.

Heitkötter, H., Hanschke, S., and Majchrzak, T. (2013). Evaluating cross-platform development approaches for mobile applications. In Cordeiro, J. and Krempels, K.-H., editors, *Web Information Systems and Technologies*, volume 140 of *Lecture Notes in Business Information Processing*, pages 120–138. Springer Berlin Heidelberg.

Kurka, D. (2013). mgwt - Making gwt Work with Mobile. http://www.m-gwt.com/.

Mittal, R., Kansal, A., and Chandra, R. (2012). Empowering developers to estimate app energy consumption. In *Proceedings of the 18th annual International Conference on Mobile Computing and Networking*, MobiCom '12, pages 317–328.

Monologue Inc. (2013). Mono framework, http://www.mono-project.com/.

Monsoon Solutions Inc. (2013). http://www.msoon.com/LabEquipment/PowerMonitor/.

MoSync Inc. (2013). MoSync http://www.mosync.com.

Motorola Solutions, Inc (2013). Rhodes http://www.motorolasolutions.com/us-en/rhomobile+suite/rhodes.

Palmieri, M., Singh, I., and Cicchetti, A. (2012). Comparison of cross-platform mobile development tools. In *16th International Conference on Intelligence in Next Generation Networks*, ICIN '12, pages 179–186.

Panacoda GmbH. (2013). The-m-project http://www.the-m-project.org/.

Pathak, A., Hu, Y. C., and Zhang, M. (2012a). Where is the energy spent inside my app?: Fine grained energy accounting on smartphones with eprof. In *Proceedings of the 7th ACM European Conference on Computer Systems*, EuroSys '12, pages 29–42.

Pathak, A., Jindal, A., Hu, Y. C., and Midkiff, S. P. (2012b). What is keeping my phone awake?: Characterizing and detecting no-sleep energy bugs in smartphone apps. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 267–280.

Raj, R. and Tolety, S. (2012). A study on approaches to build cross-platform mobile applications and criteria to select appropriate approach. In *Annual IEEE India Conference*, INDICON '12, pages 625–629.

Sencha Inc. (2013). Sencha touch, http://www.sencha.com/products/touch.

Seregon Solutions Inc. (2013). dragonrad http://dragonrad.com/.

Thompson, C., Schmidt, D. C., Turner, H. A., and White, J. (2011). Analyzing mobile application software power consumption via model-driven engineering. In Benavente-Peces, C. and Filipe, J., editors, *PECCS*, pages 101–113. SciTePress.

Yoon, C., Kim, D., Jung, W., Kang, C., and Cha, H. (2012). Appscope: Application energy metering framework for android smartphones using kernel activity monitoring. In *Proceedings of the 2012 USENIX Conference on Annual Technical Conference*, USENIX ATC'12, pages 36–36, Berkeley, CA, USA. USENIX Association.

SHORT PAPERS

Robots Collaboration based on Cloud Robotics System for Daily *Emergency Life*

Buribayeva Gulban and Taizo Miyachi
School of Information Science and Technology, Tokai University,
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
banu.b@hotmail.com, miyachi@keyaki.cc.u-tokai.ac.jp

Keywords: Cloud Robotics, Robots Collaboration, Emergency Situations, Daily Life, Emotion Recognition, Awareness.

Abstract: A person should solve unpredictable problems and evacuate from difficult emergency dangers by himself/herself. These days many different tasks can be solved by robots. However difficult tasks need more information in serious changing environments and human minds which are shared by robots and cloud servers can achieve this greater information. We propose a robot collaboration methodology in a cloud robotics computing system that can enable a robot with few resources to solve the difficult problems in both human daily life and emergency situations and provide users with useful awareness. We discuss how robots collaborate utilizing cloud robotics capabilities in order to exchange data/contexts and to find candidate solutions and create maps for emergency evacuation. We also discuss experimental results by a prototype of robot system for daily/emergency life.

1 INTRODUCTION

A person should solve unpredictable problems and evacuate from difficult emergency by himself/herself. Nowadays, robots are not only toys for children, but they are also useful tools for people of all ages in many situations. They are useful for health rehabilitation, rescue, robots for dangerous tasks, housekeeping robots and so on. Over the past few years in the field of information technology has evolved a new paradigm - cloud computing. Although cloud computing - a special way to provide computing resources has caused a revolution in the methods of providing information and services. A similar revolution in collaboration between robot and human utilizing the cloud computing that is called "Cloud robotics" (James, 2010). A robot with sensors and actuators interacts with the environment. Sensor data become useful feedback for making decisions of next effective action of the robot. This reduces the load on the computing power of the robot, as well as reduces the cost of the robot. Maps of sensor data show the locations of robots and some contexts in the serious environments. We discuss an autonomous robot in daily life as well as in emergency situations, like an earthquake, based on the cloud robotics. We also discuss experimentation by prototype of autonomous robots.

2 EMBEDDED INFORMATION IN DAILY LIFE AND EMERGENCY TIME

Inconvenient information for a person is often concealed in daily life even in ubiquitous computing society. People can acquire a limited volume of useful information in a disaster. Humans need intelligent assistants in order to remind them of important issues and solve suspended problems.

In the case of daily life, the parents should look after the child, although they are very busy and are not always able to find the change in their child's grades. Some children don't want to tell their parents about their grades or in another case some of them can just forget to tell their parents about the note from their teacher. Skipping this information will be a problem for the children's study, because parents can think that their children go to the school every day and everything is fine with them. In this daily life situation a robot will be useful for the people.

In case of disaster situation it's very vital for people to be informed to quickly evacuate from a

dangerous area. Some human nature has negative effect such as panic, "normalcy bias," "catastrophe forgetting" and this will be reason of making mistakes like wrong staying zone(not safe), not effective route to get to safe places. Some people are thinking that their place (house) is strong and it isn't necessary to be evacuated. Robots with cloud can be an intelligent assistant for humans. We discuss robot rules in order to avoid these problems by providing real-time information and to help people get to the safe please as soon as possible.

3 ROBOT COLLABORATIONS AND AUTONOMY FOR EFFECTIVE SOLUTIONS

Robot network system has two sets of collaboration frameworks in (A) daily life and (B) disaster evacuation.
A robot should acquire trust from its owners, and it should also contribute to the owners in both cases (A) and (B).
A robot has mainly six kinds of "Autonomy" for both daily life and emergency evacuation.
Daily Life. (1) Relative analysis, (2) Best solutions in each field, and (3) creation of motivation.
Emergency Evacuation. (4) Universal solutions in all approaches, (5) Embedded information, and (6) Psychological assist

3.1 Daily Life Awareness: Learning Management by Relative Analysis

A student should have good motivation and actively proceed in his/her learning. When a student gets low grades on their examination a personal robot discovers weak points of its owner, like serious results on a test and can autonomously find good reference books and tutors in the cloud servers corresponding to the level of serious grade. It will keep communication between teachers and parents open. The robot acquires the latest grades and analyzes contexts of changes of the grade. The robot also autonomously analyzes the reasons of the serious grade of the student based on the change of grades of the student's friends. Then the robot suggests these solutions to the parents:

If the result of the examination is worse than 1-30 %, it recommends a list of textbooks for improvement to help the student prepare for the next exam. The robot makes a request list of textbooks from cloud robotics.

If the result is worse in the 31-50 % that gives a list of tutors number 1, with a score worse than 51-100 % is another solution that can be given to parents by the robot to improve the test results of the child.

The parent can consult with the student utilizing candidate solutions. Daily assessment of students, teacher's notifications, school events will be saved in the cloud robotics. Related with the student achievements the robot will give advice to parents by analyzed data. It will be easier for parents to be informed about their children at the school. (see Fig 1).

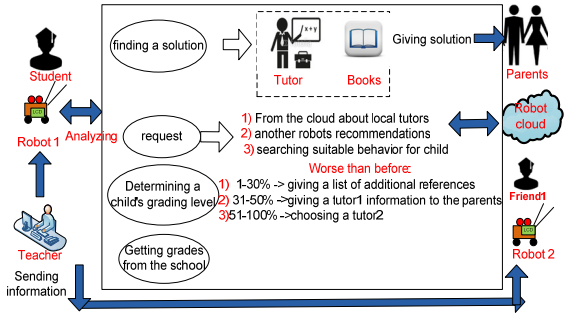


Figure 1: Learning management by robot collaboration.

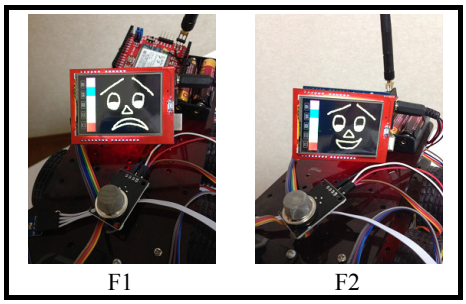


Figure 2: 80 faces of robot. Emotions such as F1 (joy 1.0 and sadness 0.4) and F2 (joy 0.4 and sadness 1.0).

3.2 Emergency Awareness Avoiding Danger

3.2.1 Communication between Robots and the Cloud by using Clone-based Model

Flexible adaption against sudden attack of disaster need autonomous actions of each robot. We propose "Clone-Based Model" for collaboration between Robots and cloud with redundant servers (see Fig 3).

"Clone-Based Model: each robot has a corresponding system- level clone in the cloud. A task can be executed in the robot or in its clone. The set of robotic clones also form a peer-to-peer network with better connectivity than the physical

ad-hoc M2M network. Moreover, this model allows for sporadic outage in the physical M2M network" (Guoqiang, 2012).

Robots and cloud share the data and history of the data. The shared data enables them to analyze Big Data.

For example, serevers in Tokyo would be broken by a disaster. Then the robots can send and receive information from the server in Osaka . Parallel backup data is necessary because we can not imagine the destructive force of the earthquake . Robots R1 (west area), R2 (east area), R3 (north area), R4 (east area) are in the local area and should autonomously detect risks and share them. (see Fig 3).

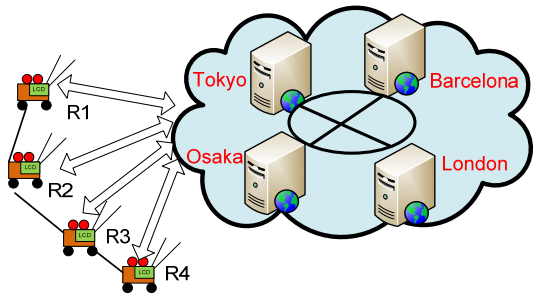


Figure 3: Clone-Based Model of robot collaboration system.

(rc1) All robots will upload new sensors data to the cloud. In parallel all servers(clones) will get new sensors data from the robots.
(rc2) The cloud will collect real- time photos of situatuions and can send these to the robots.
(rc3) The robotic cloud computing will make an evacuation map in accuracy by using data from the J-SHIS and the robots.
(rc4) By using updated data the robotic cloud computing can prepare an optimal route for each robot's owner.

Robot and cloud robotics have reliable management functions of change of context in each field. It has three historical types of data: "old", "current" and "new"(see Table 1). In a disconnection situation case the robot gets new data at its current place from the cloud it also takes the latest data of the other 100 related objects from the cloud robotics.

Table 1: Data states during disconnection situation.

	robot	cloud
Old data	1	1
Current data	1	1
New data	0	1

(1) Robot Assists based on Cloud Information. Evacuation Guide. Robot uses two types of earthquake sensors. A swing sensor in its body and P-wave sensor attached to a wall or fixed pole. P-wave sensor detects the first wave (P-waves) of an earthquake. This p-wave sensor data from "Quake Alarm" device will send to the cloud by WiFi modems. The robot responds to the previous earthquake compression waves which are inaudible to humans because the frequency of the wave is less than 20Hz, just animals can react to this type of wave. The robot analyze two kinds of sensor data by inner swing sensor, P-wave sensors, and general data from J-SHIS, and make a quick decisions for evacuation's actions from a wide area point of view.

J-SHIS can provide a general evacuation map. In case robot they can give information about real-time safety places. These two maps will be different. Because in general evacuation map they are 5 nearest safety places with routs, but some places can already damaged and the robot "A" will send this damage to the cloud. By comparing two maps the cloud can make decision and provide evacuation map with effective routs. Also most of residents may not know second safety place. The robot will be used as an evacuation guide with evacuation map.
Most Japanese people think it is not necessary to evacuated before/after first wave of earthquake since Japanese houses are usually strong for seismic activity. They stayed in the houses.

Rules of robots during earthquake:

I. Providing information about the first wave and the next wave of earthquake
II. Reducing psychological forgetting (panic) of people
III. Making evacuation map and effective route to safety place by using big data from the robotic cloud.
IV. Providing real-time information about local area to people and J-SHIS.

(2) Normalcy Bias. In The Great East Japan earthquake about 70 % of residents did not evacuate although they knew the tsunami warning of the government in text and reading out by an announcer. That is why only giving warning of tsunamis is not enough since the citizens have been safe for 38 years. They believe it would be enough that only playing no serious damage by tsunamis. They need to acquire the motivation and chance to start evacuation from the giant tsunami.

Visual information allows residents to quickly understand serious situations and to make the decision to leave their houses. Such photos/video (see Fig. 6) of running to the safe place enables residents to avoid psychological disorders such as

normalcy bias and catastrophe forgetting (Miyachi, 2013) and quickly start evacuation in the safe direction. We propose 5 seconds duration video (photos) that show evacuation actions by people who already have evacuated in real-time by the robot. To avoid this kind of psychological problem the robot can change faces corresponding to the degree of an impending earthquake (see Fig. 2) (Schmitter, 2008) and also showing real-time pictures from another robots about people evacuation.

(3) Contribution to Japanese Earthquakes. Earthquakes usually cause fire, since Japanese houses are built by wood and there are many restaurants in the towns in a same street. If strong winds expand the fire over the first shelter, residents should change the place of refuge from the first safe place to the second safe place and know safe routes to get there. Most residents do not know such information. The robot should show the residents the second safe place and the safe route while sensors of robot are detecting dangers for safe mobility to the safe place. Fallen concrete block, fire in the old wooden houses, road collations, and evacuation ability of aged persons should be provided by the robot and knowledge in the cloud servers.

(4) Contribution to Kazakhstan Earthquakes. There are a lot of industrial companies in Almaty city (see Fig. 4). Each of them can become flammable zones (green icons-flammable zones and blue icons-safety places, yellow icon -person's house). In case of a person's house being near the industry area the robot would recommend to go to the place "A0" considering the strong north wind. Robot could find a safe direction and safe places to evacuate based on the direction of the wind from server of "National Hydrometeorological Service of the Republic of Kazakhstan". Robots should have "decision table for safe direction/safe place" in order to quickly find them. The decision table becomes an explanation of why the decision would be the best solution.

3.2.2 Collaboration between Robots without Cloud

This communication allows robots connect with each other without cloud access point by wireless communication, such as WiFi module or WiMax etc. The Robot R1 is near dangerous are (fire,gas etc) R1 will send a request about area to the friends robot R2 (neareast robot). If the R2 will be near not safety place , the R1 will recomendate to an owner to leave area near R1 and R2.



Figure 4: Google map. Almaty city with flammable zones.

(rr1) A robot autonomously gets sensor data and detects danger.
(rr2) A robot communicates with a robot near it and checks whether the other robot has similar situation or not when it detected the danger.
(rr3) A robot also sends the other robots awareness of such danger that it detected.
(rr4) A robot shows its owner awareness with the information from the robot that detected the danger.
Data exchanging between robots without the cloud is should be simple to save memory and computation resources. Robot R1 will get data from robot R2 in a table. The table includes information: which robot (ROBOT_ID), in which area (AREA_ID), has what kind of situation (earthquake, tsunami, fire etc) (SITUATION_ID), when (TIME), what is data (DATA) and situation level (DANGER_LEVEL).

A good point of this communication is that it enables robots to autonomously find specific areas and minimal additional computation and memory resources.

4 EXPERIMENTS

Experiment 1. Learning management by Student's Grade. The robot found good reference books and a tutor for the weak points of the student based on collaboration with friend's robots through the cloud robotics system. The robot could recommend the parent the best reference books and the tutor for characters of the child.
Subjects: International Students (between 24 and 33 years old)

Question1. "What kinds of recommendations (A: books, B: tutors, C; books and tutors) do you give for each cases of students to improve their grades?"
Case 1: one student's grade decreased and the other student's grades did not change
Case 2: Half of the student's grades decreased and other student's grades did not change
Case 3: All students' grades decreased.

We found that all international students expect to get tutors and books in case 3. In case 2 we see that about 70% of subjects needed good tutors although half of subjects decreased the grades. We could make sure that the awareness with candidate of tutors for parent would become useful in case of both case 3 and case 2. International students might be very sensitive for decrease of the degree. More than 63% of the students can study by themselves in case 1.

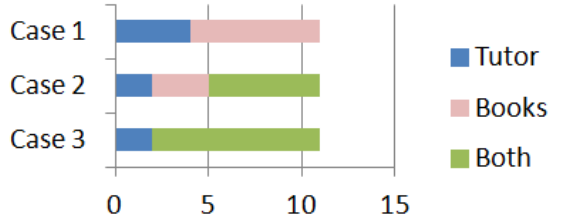


Figure 5: Preferred awareness for the level of grade.

Experiment 2. Avoiding Psychological Disorders

As we discussed in section 3 about providing real-time video (photos) with name of area by different areas robots it will be shown on a TFT LCD as in below figure 6.



Figure 6: Simulation showing photos in case evacuation.

Experiment 3. Autonomous robot detection based on crone model. The types of sensors are fire, gas, rain/snow, ultrasonic, and shaking (see Fig 7).

The robot could detect fire, gas, rain/snow, ultrasonic, and shaking (see Fig. 8).

All sensors (see Fig 8) were attached on the Arduino Uno board. Arduino is a hardware computing platform, the main components are simple input-output board and a development environment for language Processing/ Wiring.

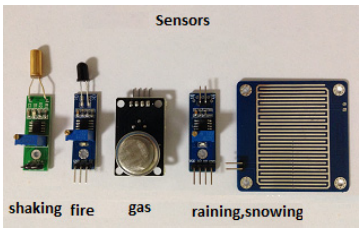


Figure 7: Sensors.

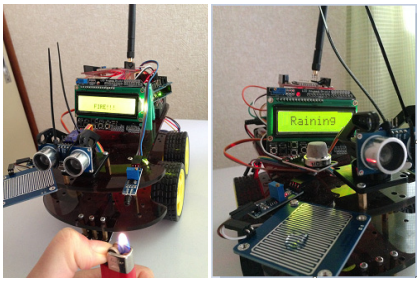


Figure 8: Detecting fire and rain .

Programming language is C. Initial step of experiment was with fire, raining sensors. Fire detection is necessary for emergency situations to avoid flammable zones. Robot showed messages such as "No fire"/"Fire". Rain detection is also useful to analyze weather in case of fire situation can be changed. "No rain"/ "Raining" on the LCD display corresponding to the rain detection.

Experiment 4. Emergent Awareness between Robots. We experimented communication (data exchanging) between the robot and computer with fire sensor through Bluetooth channel. We used for the robot a "Slave HC-06 Wireless Bluetooth RF" module to send sensor data to a computer terminal and for the computer "Tera Term" a terminal emulator. When the robot could detected fire (see Fig 8) on the Tera Term we see text "Fire area" in case detecting "no fire" on the terminal we see the text "Safety" (see Fig 9). In the next step a Bluetooth module will be replaced with WiFi module also instead of a terminal emulator we will use a web server.

5 CONCLUSIONS

In this paper we propose robots collaboration system based on cloud robotics that provides useful awareness for unpredictable problems. We discuss how robot collaboration works in both daily life and emergency life. We also discuss six kinds of robot autonomy and awareness based on relative analysis.

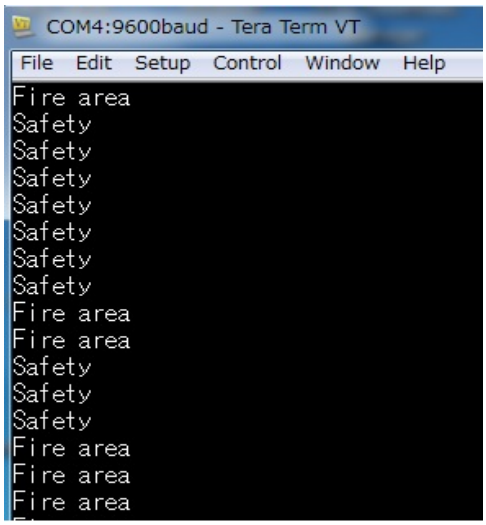


Figure 9: Tera Term is getting signals from the robot's fire sensor through Bluetooth channel.

We made a prototype of a robot system. By detecting fire/dangerous gases and sending these signals to the computer via Bluetooth channel. In addition, we did simulations by creating various faces for the robot that can be used to express the robot's emotions depending on the situations and changing real-time photos on TFT LCD to avoid human psychological disorders.

Our next step will be by using "p wave" sensor to detect the first wave of earthquake and sending this signal to the server by using SainSmart WiFi Wireless Shield WizFi210. Also we will make a second similar robot to test collaboration with each other. Then we will check how the robot will make an evacuation map by using data about the first wave of earthquake and signals from other sensors like radiation etc. from the server. Also we are going to use emotion recognition with Kinect and voice recognition system that the robot can understand voice commands.

REFERENCES

Ben Kehoe, Akihiro Matsukawa, Sal Candido, James Kuffner and Ken Goldberg, 2013. *Cloud-Based Robot Grasping with the Google Object Recognition Engine*. *IEEE International Conference on Robotics and Automation(ICRA)*.

Caroline Schmitter and Ryosuke Masuda, 2008. *Research on multiple facial expression of robots*, *Journal of the School of Information Science and Technology*. Tokai university. Vol XXXIII.

Dominique Hunziker, Mohanarajah Gajamohan, Markus Waibel and Raffaello D'Andrea, 2013. *Rapyuta: The RoboEarth Cloud Engine*, http:// roboearth.org/uploads/RCE2013.pdf.

Guoqiang Hu, Wee Peng Tay, and Yonggang Wen,2012. Cloud Robotics: Architecture, Challenges and Applications. *IEEE Network* May/June 2012.

James J. Kuffner, 2010. *Cloud-Enabled Robots*, *IEEE-RAS International Conference on Humanoid Robotics*.

LaBar KS, Disterhoft JF ,1998. *Conditioning, awareness, and the hippocampus*. Hippocampus 8 (6).

McCloskey, M. and Cohen, N.,1989.*Catastrophic interference in connectionist networks: The sequential learning problem*. in G. H. Bower (ed.) The Psychology of Learning and Motivation: Vol. 24, 109-164, NY: Academic Press.

Miyachi T., G.Buribayeva, Saiko Iga and Takashi Furuhata, 2013. *Evacuation Assist from a Secquence of Disaster by Robot with Disaster Mind.*

Provision of Personalized Data via Mobile Web Services in eHealth Scenarios

Marc Jansen[1,2], Abbas Siddiqui[1] and Oliver Koch[1]
[1]*Computer Science Institute, University of Applied Sciences Ruhr West, Bottrop, Germany*
[2]*Department of Media Technology, Linnaeus University, Växjö, Sweden*
{marc.jansen, abbas.siddiqui, oliver.koch}@hs-ruhrwest.de

Keywords: Personalized Data, Context-awareness, Service Provision, Service Consumption, Middle-ware, eHealth, mHealth.

Abstract: In societies where the demographic change leads to a more and more unbalanced state between the elderly and all other different age groups, there, health management becomes one of the most significant problems, e.g., in order to allow what is often called successful aging. Ubiquitous use of smart mobile devices enables gathering of personalized data enriched with contextual information. This information can ideally be used in modern eHealth scenarios, resulting in mobile health (mHealth) scenarios. This paper describes how highly personalized and sensitive information, such as vital signs enriched with contextual information of a patient, can be stored at mobile devices and provided via modern web technologies for later-on analysis and health monitoring. Therefore, this paper presents a number of different scenarios in which such mobile technology provides certain benefits, discusses benefits, drawbacks and challenges of such an approach and describes an example implementation of an mHealth scenario.

1 INTRODUCTION

Nowadays, smartphones and other mobile devices are used tremendously and this trend is still growing rapidly. Here, the functions of smartphones in use are much more than simple voice calling and text messaging. The mobile users use mobile phones to access their mail accounts, to share their calendars or to even use their devices as a routing device (i.e., the ones equipped with the GPS). Therefore, smart mobile devices collect various information about their users such as personalized information, personal/business contacts, messages (emails, short text messages like SMS, …), social interaction and social contacts, calling history, work history and footprints of internet usage (e.g., the browsing history of a mobile web browser). Additionally, the set of applications, a user has installed on his device, provide a rough profile of usage of the device itself.

Two major advantages from the increased usage of mobile devices are: On the one hand, the devices are mobile and could therefore ideally be used to follow their users round the clock and ubiquitously. And on the other hand, a modern internet connection is provided with most of the smart mobile devices, this makes such kind of mobile devices a powerful and valuable tool. The second reason for the success of smart mobile devices is, that these devices are usually equipped with a large set of different sensors, allowing to contextualize the users' current tasks, e.g., by using the GPS sensor of a modern mobile device, the current position of a user could also be determined accurately, so that additional information about the task could be gathered, either to provide better support to users during the task or to store additional contextualization data for later analysis. In addition to GPS sensors, other sensors like acceleration sensor, motion sensors, a digital compass and alike, could be used on modern mobile devices. In the context of eHealth scenarios, additional sensors for monitoring current vital signs of the user could be connected to a mobile device.

The idea presented in this paper is twofold, on the one hand an architecture is presented that allows to store the data on ones' own device and not in some central database. Here we argue that this is a big advantage (in contrast to modern Cloud Computing and Web 2.0 scenarios in which more and more data is stored centrally), having in mind that health data is always personalized data with a big demand for security. The second idea is provided in this paper is to combine vital data (gathered by

specialized eHealth sensors connected to a standard mobile device) and contextualization data (gathered by standard sensors of a mobile device) and to provide it in an authentic and secure way, which without the usage of mobile devices could hardly be achieved. Additionally, two scenarios are presented in detail to demonstrate the use of that kind of technology. For one of the scenarios, an already existing implementation is described. Here, both on the implemented middleware and on the client side, web based technologies have been used in order to provide a user friendly implementation.

The rest of this paper is organized as follows: first we provide an overview of related work in the next section. After this we describe our approach in section three, followed by a discussion of the presented approach in section four. Section five describes two scenarios in which the presented approach provides benefits and section six describes the web based implementation of one of these scenarios. Finally, section seven provides a conclusion and an outlook to future work.

2 RELATED WORK

Providing Web Services on mobile devices was probably presented first by IBM (McFadden, et al., 2003). In this work a solution is presented for a specific scenario in which Web Services are hosted on mobile devices. A more general approach for providing Web Services on mobile devices is presented in (Srirama, et al., 2006 and AlShawan, et al., 2010). In (Li and Chou, 2011) the authors suggest a different approach, focusing on the optimization of the HTTP protocol for mobile Web Services provisioning. Importantly, none of the mentioned approaches manages to overcome certain limitations of mobile devices, e.g., permanently changing networks, IP addresses from networks with network address translation (NAT) or the fact that mobile devices are usually not designed to be always online (might be switched off, might have not network connection, …).

An additional approach that manages to overcome these problems, is presented in (Jansen, 2013a). This approach utilizes a central proxy infrastructure, that allows on the one hand to tackle the mentioned challenges and on the other hand provides a stable infrastructure for mobile device to provide standardized Web Services. Furthermore, the central proxy provides the major part of the technical infrastructure that allows to use standardized protocols for the deployed Web Service at the mobile devices.

Additionally, the work presented in (Jansen, 2013b) argues for a new perspective to Web Services especially if those services are deployed to mobile devices. Here, the author argues that one of the major benefits of mobile devices is the fact that they are mobile. In connection especially to mHealth scenarios, this fact allows to gather different kinds of data from the sensors of a mobile devices that allows to easily contextualize the user (in most mHealth scenarios, the patient) in his/her current task/situation.

3 APPROACH AND DISCUSSION

The approach we used for the implementation of the example scenarios is based on a standardized Web Services. As already explained, a number of different approaches exist that allow the provision of Web Services on mobile devices. Since from our point of view, the approach described in (Jansen, 2013a), solves the most pressuring problems, we decided to base our implementation on this approach.

In addition to this, we implemented a middleware layer allowing a loose coupling between the provider of the mobile Web Services and the consumer service in our case a web-based application that visualizes the data measured by different sensors and in different training sessions.

Another advantage of the approach described in (Jansen, 2013a) is that it allows to store requests at the central proxy, e.g., if the device that provides the mobile Web Service is not connected to a network. As soon as the device is connected again, the request could be performed by the device and the result is available for the client.

Additionally, this approach implements Web Services for mobile devices based on standardized protocols, so that developers who are familiar in using Web Services do not need to learn a whole new technology, but can rely on knowledge they already have. Furthermore, also on the mobile Web Service provider side, the approach utilizes annotations pretty much in the same way as for the implementation of Web Service on non-mobile devices, providing again a very flat learning curve for the new technology.

The chosen approach provides a number of benefits, e.g., since the measured eHealth data only remains on the mobile device of the patient, there is no need for a central server infrastructure. This results in lower management costs, easier to achieve

security and no additional costs for central backups. Additionally, there is no need for central personnel, space, hardware and software provisioning and no central costs for energy.

On the other hand, this de-centralized approach also provides some drawbacks. Since the data is not stored centrally, hence, no central backup strategy is applicable, but the patient him-/herself is responsible for his/her data. Also security management would be easier to accomplish in an architecture with central data storage. Additionally, new questions with respect to the security of the data come into play while the data remains on mobile devices: what happens if the device gets lost or stolen? Therefore, of course, the personal data should be stored encrypted on the device of the patient. Still, encryption would just solve the problem of privacy (no one else but the owner of the mobile device would be able to access the data), but not the problem of availability of the data. Once the device is no longer in reach of its owner, the data would no longer be accessible for future analysis.

Besides the challenges that come from the described drawbacks, an additional challenge comes into play considering the target group for this kind of mHealth applications. The usual target group for this kind of applications consists of elderly. Since this target group is usually not specifically well used to handle mobile devices, a clear and precise explanation of the necessary steps and the used technology is more than ever necessary.

4 TWO EHEALTH SCENARIOS

Key strengths of mHealth system architectures based on mobile Web Services encompass the following aspects:

1. On-demand transmission of data to Health professionals: the doctor receives only the data that really interests him and especially at the point in time he is interested in.
2. Data remains in the direct disposal area of the patient: the need for informational self-determination is thus optimally fulfilled.

The following two scenarios, describe the use of Web Service based mHealth applications by illustrating the two aforementioned strengths.

4.1 Scenario 1: Surgical Preparation

The central source of revenue for hospitals is the surgical area. This is the actual place where the decision is taken whether a hospital makes profit or deficit. Therefore, usually all hospital processes are aligned to the optimal utilization of the surgical area. Any delay in the process or even surgeries that need to be canceled or postponed at short notice, cause a more or less considerable economic loss.

Prior to surgery, patients usually must follow certain rules of conduct and take or discontinue certain medicine for several days or weeks. Has he/she not done so (e.g., medication for blood clotting was not discontinued), the sugery has to be canceled at a short notice. This may lead to an underutilization of the surgical area and consequently the associated economic losses.

A 78-year-old patient is to receive a new right hip joint. She suffers from a lighter hypertension and type II diabetes. At the same time she tends to water retention, which manifests in an abrupt increase in her weight. Normally hip replacement operations rank among "standard" operations with less risk involved. But, because of the patients' secondary diagnoses, however, there is a significant surgical risk. Prior to the operation her risk factors must therefore be closely monitored. To shorten the period of rehabilitation after the surgery, the patient should move as much as possible in advance to the surgery and complete a specific exercise program to strengthen certain muscle groups.

For several months already, the patient has a smartphone. At her last doctor's appointment, four weeks before the surgery, a mobile mHealth application for collecting and managing various sensor data (in this case, blood pressure meter, blood glucose meter and scale connected via bluetooth), is installed and configured on her mobile phone. The data is provided via a mobile Web Service, which is also installed on the device. At the same time the female patient receives an introduction into a sensor package that will be handed out to her. Part of the mHealth application are also information films, in which the preliminary physical exercises are provided to her and she gets also informed about the risks and the progress of the medical intervention.

In the following four weeks, the female patient regularly measures her blood pressure, blood sugar and weight. This data is automatically enriched with location, time, and motion information from her mobile phones sensors (GPS, accelerometer, proximity sensor, etc.), allowing to contextualize the data gathered by the medical sensors.

Thus, noticeable measured values (e.g. increased heart rate and blood pressure due to sports activities) can be interpreted, according to the context in which they are occurred.

The data is generated using the internal sensors of smart phones and possibly even enriched by data from external web services (e.g. historical weather data). By using a feedback function, the patient may give feedback, for example how she felt during the exercise and the measurements. The logged feedback and sensor data can be queried by the hospital on-demand. When retrieving the data the hospital doctors can decide if they want to get all data or only aggregated data in terms of a status report (surgery possible / not possible). In addition, the mHealth application detects emergency situations (e.g. very high blood sugar levels) and alerts the hospital if necessary. The surgical management can check at any time whether the surgery will be possible, a change in behavior of the patient is necessary or an adaptation of the surgical plan is required.

Even at the stage of rehabilitation the Web Service on the patient's smartphone can provide on-demand information about the patient compliance (e.g., lack of physical exercise), allowing the timely intervention of the treating physicians.

4.2 Scenario 2: Regular Data Measurement by Patients

In the second scenario we imagine a 33 years old male patient diagnosed of having a genetic disease that disallows his body to effectively fight against blood lipids. Therefore, he is likely to have an increased risk of an attack. In order to lower this risk, besides a special medication, the patient is asked to have 24 hours blood pressure measurements at least once a year. A couple of years ago, this was more or less unhappy day every year, in which the patient had to wear a heavy blood pressure monitor over a whole day. Lately, his doctor told the patient that with new upcoming technology in the mHealth sector, the process for 24 hours blood pressure monitoring could be eased a lot.

By connecting a small and light blood pressure sensor to his smartphone, the patient is easily able to collect the necessary data in a convenient way. Furthermore, by collecting additional data from the mobile device of the patient, e.g., the current geolocation and/or the acceleration at which the patient is currently moving, the gathered data could be contextualized in order to ease the task for later analysis.

Happy to know about these new possibilities, the patient is willing to take his next 24 hours blood pressure measurements with this new technology. Two weeks after the measurement was done, the patient has a new appointment with his doctor in which the data from the measurement should be discussed. During this discussion, the doctor recognizes a tremendous increase in the blood pressure at 17:03 at the day of the measurement. First, neither the patient nor the doctor could explain this increase, but by looking at the contextual data gathered along the blood pressure values, the reason for the increase of the blood pressure became clear. By analyzing the geoposition of the patient and his acceleration at the time of the increasing blood pressure, it turned out that the patient was watching his famous soccer team playing against another local team. Knowing this, the patient remembers that at this exact time, the opponents of his famous team managed to shoot a goal. Being upset about this event clearly explains the increase in the patients' blood pressure values.

Again here in this scenario, beside the two major advantages of on the one hand privacy of the data (the data remains with the patient and is not stored centrally) and the just in time provision of the interesting data to the right person (in this case the doctor) could easily be seen. Furthermore, advantages of using modern mobile devices in mHealth scenarios, like the rich source for the contextualization of the patient are also obvious.

5 IMPLEMANTATION OF AN EXAMPLE SCENARIO

The prototype consists of two parts namely Web Services (i.e., running on a mobile device) and a web based client (i.e., running in a browser).

The following paragraphs describe the detail about these aforementioned components of the prototype followed by the interaction among the components.

5.1 Mobile Application

Android based mobiles are selected for the implementation of the concept as according to the International Data Corporation (IDC) Worldwide Quarterly Mobile Phone Tracker's 2013 report, android based mobiles are the market leaders. An Android based application is developed to collect the various sensors data.

There are two categories of sensors, the ones integrated in a mobile phone and others which are externally linked via well-known communication technologies (e.g., Bluetooth, Wi-Fi).

In this particular implementation, the location of

a patient is determined by an integrated global positioning system (GPS), and the speed of movement is measured by a built-in accelerometer, which measures the force of acceleration caused by the gravity or weight to determine the speed. The standard Android library is used to retrieve the sensors data.

Besides the integrated sensors, to measure the heart-rate and, to measure the instant speed of a patient, a Zephyr sensor (Zephyr HxM BT – Heart Rate Monitor) is used. The sensor is mounted to a smart fabric based strap, which must be worn around the chest. The data is periodically collected via Bluetooth. The HxM sensor provides a Bluetooth API to collect the data in a specific message format.

After the collection of the data from all those aforementioned sensors, the data is saved by the application shown in Figure 1 to a SQLite database.

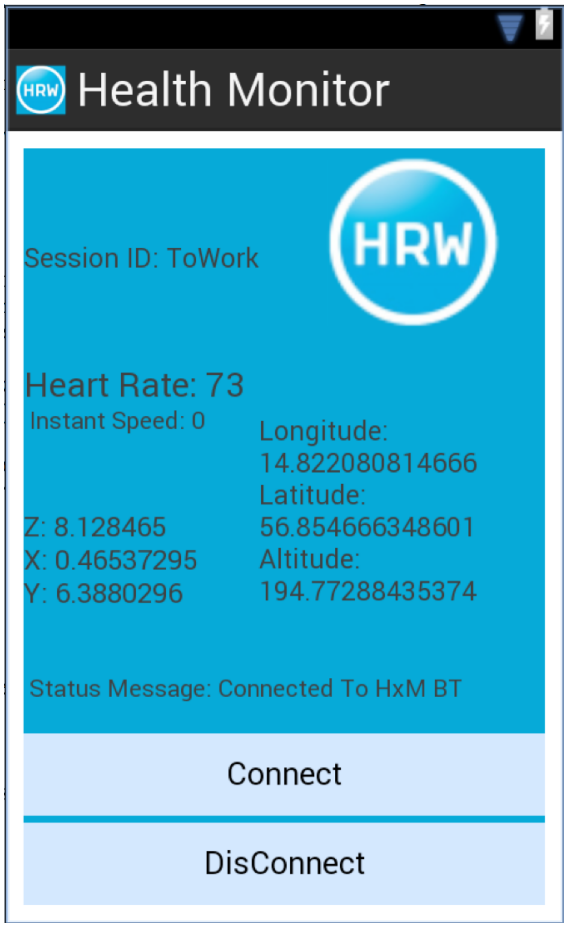


Figure 1: A screenshot of the app collecting data from an mHealth session.

A SQLite database runs without a server, which makes it a suitable choice for the energy limited mobile devices. Nevertheless, SQLite data is very application related and dependent, once an application is uninstalled then the data will no longer be available. The database contains two tables "sessions" and "sensorsData".

Besides retrieval and storage of the sensors data, another part of the application handles the incoming request as a mobile Web Service. It fulfills the authentic incoming request by sending the stored data on the device to the client.

5.2 Web Client

The client runs in a browser and can be invoked from everywhere. Most of the functionality is implemented in JavaScript and HTML5. Nevertheless, due to the same origin restriction, some façade, according to the façade/proxy design pattern (Gamma, et al., 1995) services needed to be implemented on the server side. For this implementation we have chosen PHP.

After a successful authentication (e.g., as a doctor), the web application sends a Web Service request to the mobile device of the patient in order to receive a list of all available sessions. By choosing a certain session, another Web Service request is send to the mobile device that receives the data gathered during this session. This data could then be visualized in a tabular format or as a representation in a Google Map, where, e.g., the data is located at the correct geoposition and different parts of the data could be visualized according to the geoposition at which they appeared. An example of the visual representation on the client side in a Google Map could be seen in Figure 2.

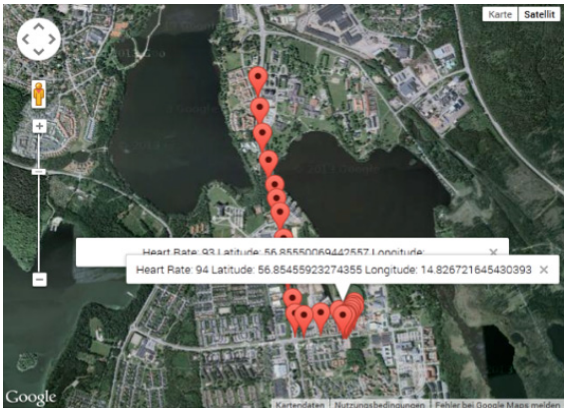


Figure 2: Visual representation of gathered health data in a Google Map.

5.3 How All Work Together

The functionality of the prototype is described from

the two different perspectives namely a user (e.g., a patient), and a user of the collected data such as a general physician, a family member or patient himself.

From the mobile-health-moniters' user (e.g., a patient) perspective, a user starts the application which collects the sensors information on his/her mobile device. Before starting a new session to record the sensors data, a user must wear a HxM chest strap with mounted heart-beat sensor to collect the real heart-beat of a user, otherwise a default value (i.e., which is "0" in this case) will be stored. Once the session is started, every few seconds the new values from the location sensor, 3-axis accelerometer and HxM are retrieved and stored in the "sensorsData table in the "mhealth" database. Let's assume, a user started a session during the workout, once the workout is over, the user decides to discontinue the session and clicks the disconnect button, once the session is disconnected a new entry in the "sessions" table will be stored with start-time and end-time of the session.

When a concerned person (e.g., a personal doctor, a family member, a care take, or the user/patient him-/herself), let's assume, a house-doctor wants to know about the patient who is using the mobile-health-monitoring app. The house-doctor will enter the web-link in a browser to invoke the mobile Web Service running on the mobile device. Before accessing the patients' data, the house-doctor must identify him-/herself with credentials through a username and a password. After the house-doctor is identified, all the stored sessions on the patients' mobile device are retrieved via the mobile Web Serivce. Now, the house-doctor can select any session to analyse the measured data.

6 CONCLUSIONS AND FUTURE WORK

Wrapping up, this paper provides a discussion about the usage of standardized Web Services on mobile devices like smartphones in the context of mHealth scenarios. We provided some example scenarios, a discussion about the current state of the art and also about benefits and drawbacks of the presented approach. Last but not least, a web-based implementation of one of the example scenarios was presented.

Nevertheless, beside the obvious benefits of such an approach, there are also some topics that need to be considered much more intensively in future research. One of the major things to have in mind, is the security of the data on the phone of a patient. More likely, data security (especially data privacy) could be much easier realized for a central database (i.e., administrated by IT specialists). On the other hand, data security is anyway a topic that users of smart devices have to deal with and there are already quite powerful technologies available that increase the level of security for mobile data. Here, an important task for future research efforts will be to integrate such technologies in the described scenarios.

Another aspect comes into play if we consider that the different kinds of data possibly gathered by mobile devices need to be arranged and visualized in a way that allows doctors to easily interpret this data. Since doctors are usually not specialists in data analysis, additional efforts need to be considered for the presentation and analysis of the data.

Last but not least, the reliability of the medical data gathered by not necessarily medical devices is an issue that needs to be tackled.

REFERENCES

F. AlShahwan, K. Moessner, "Providing SOAP Web Services and REST Web Services from Mobile Hosts", In: *Fifth International Conference on Internet and Web Applications and Services (ICIW 2010)*, pp. 174-179.

E. Gamma, R. Helm, R. Johnson, J. Vlissides, Design Pattern – Elements of Reusable Object-Oriented Software, Addison-Wesley. 1995.

M. Jansen. Analysis and Improvement of Energy Consumption for Providing Mobile Web Services. In.*International Journal of Soft Computing and Software Engineering*, DOI: 10.7321/jscse. 2013a.

M. Jansen. About the Necessity to Change the Perspective for Mobile Web Services. In: Proceedings of the 15th *IEEE International Symposium on Web Systems Evolution*, 2013b.

L. Li, W. Chou, "COFOCUS – Compact and Expanded Restful Services for Mobile Environments", In: *Proceedings of the 7th International Conference on Web Information Systems and Technologies*, pp. 51-60, Noordwijkerhout, The Netherlands, 2011.

S. McFaddin, C. Narayanaswami, M. Raghunath, "Web Services on Mobile Devices – Implementation and Experience", In *Proceedings of the 5th IEEE Workshop on Mobile Computing Systems & Applications (WMCSA'03)*, pp. 100-109, Monterey, CA.

S. Srirama, M. Jarke, W. Prinz, "Mobile Web Service Provisioning", In: *Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006)*, p. 120, Guadeloupe, French Caribbean.

AUTHOR INDEX

Aburub, F. 310
Adamkó, A. 226
Andreica, A. 214
Arenas, H. 152
Arroyo-Figueroa, G. 332
Baumann, A. 361, 369
Behrend, A. 143
Beul-Leusmann, S. 411
Biancalana, C. 70
Blazejczak, A. 257
Botto, M. 208
Boufaida, M. 318
Caballero-Gil, P. 238
Caldeira, M. 375
Carvalho, H. 375
Castejon, H. 185
Cavalli, A. 168
Cavendish, D. 200
Chakhar, S. 50
Chiru, C. 294
Chroni, M. 177
Ciman, M. 423
Colin, J. 17
Connolly, J. 381
Connolly, R. 381
Corbin, T. 27
Cruz, C. 152
Czibulla, G. 214
Decker, S. 119
Erbaru, A. 294
Escalona, M. 135
Fabian, B. 361, 369
Fakhfakh, S. 220
Fidjestøl, A. 185
Finn, S. 276
Fränti, P. 62
Fylakis, A. 177
Gaggi, O. 423
Gasparetti, F. 70
Gomes, J. 375
González-Castro, S. 332
González-Huerta, J. 208
Gui, H. 193
Gulban, B. 435
Gulla, J. 185
Harbelot, B. 152
Hausenblas, M. 119
Hernández, Y. 332
Hernández-Goya, C. 238
Hoang, M. 17
Holgersson, J. 269
Insfran, E. 208
Ishizaka, A. 50
Jakobs, E. 411
Jansen, M. 441
Järvinen, H. 93
Käbisch, S. 40
Karagoz, P. 160
Karnstedt, M. 119
Kermani, M. 318
Koch, O. 441
Koide, H. 200
Kollár, L. 226
Krempels, K. 411
Küng, J. 214
Kuntschke, R. 40
Kushik, N. 168
Labib, A. 50
Lamotte, W. 257
Li, X. 354
Lian, W. 354
Limongelli, C. 338
Lischke, M. 369
Loula, P. 127
Mackie, L. 400
Madlberger, M. 247
Mahdi, W. 220
Malgosa-sanahuja, J. 285
Mallouli, W. 168
Manthey, R. 143
Manzanares-Lopez, P. 285
Mariescu-Istodor, R. 62
Marsico, M. 338
McCafferty, K. 233
Mejías, M. 135
Metaxas, P. 276
Micarelli, A. 70
Miyachi, T. 435
Miziołek, J. 27
Montazemi, A. 324

AUTHOR INDEX (CONT.)

Müldner, T. 27
Muñoz-Gea, J. 285
Mustafaraj, E. 276
Nikolopoulos, S. 177
Obidallah, W. 346
O'Connor, M. 193
Oliveira, T. 387
Paviot-Adet, E. 318
Pokhrel, J. 168
Porres, I. 104
Qi, B. 400
Quax, P. 257
Raahemi, B. 346
Rakhmawati, N. 119
Rauf, I. 104
Rebedea, T. 294
Rivero-García, A. 238
Roantree, M. 193
Romão, M. 375
Rue, E. 233
Sacarea, C. 214
Saeidi, R. 62
Samsel, C. 411
Sansonetti, G. 70
Santos-González, I. 238
Saremi, H. 324
Sarrafzadeh, A. 354
Schüller, G. 143
Sciarrone, F. 338
Sedeño, J. 135
Serrano, M. 5
Siavashi, F. 104
Siddiqui, A. 441
Söderström, E. 269
Sterbini, A. 338
Stranacher, K. 82
Su, X. 185
Tabarcea, A. 62
Tavares, J. 387
Temperini, M. 338
Terhorst, L. 233
Tmar, M. 220
Torrecilla-Salinas, C. 135
Truscan, D. 104
Turk, Ž. 394
Uratani, Y. 200
Utku, S. 160
Vahdati, S. 143
Vihervaara, J. 127
Vuorimaa, P. 93
Wang, B. 193
Wiederhold, M. 411
Wijnants, M. 257
Yevtushenko, N. 168
Zefferer, T. 82, 302
Ziefle, M. 411
Zwattendorfer, B. 82, 302

PROCEEDINGS OF WEBIST 2014 | VOLUME 1

10th International Conference on Web Information Systems and Technologies

ISBN: 978-989-758-023-9 | www.webist.org

Copyright © 2014 SCITEPRESS

Science and Technology Publications

All Rights Reserved

INSTICC is member of:

Logistics Partner:

Proceedings will be submitted for indexation by:

THOMSON REUTERS CONFERENCE PROCEEDINGS CITATION INDEX