

Toward a Social Sensor Based Framework for Intelligent Transportation

Raymond Y.K. Lau
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Hong Kong SAR
Email: raylau@cityu.edu.hk

Abstract—Intelligent Transportation Systems (ITSs) are supposed to constantly provide drivers with useful navigation information (e.g., the shortest driving paths, real-time traffic conditions, and other road condition data) by collecting raw data from physical sensors and various media. Previous work mainly considers how to develop ITS frameworks and how to extract useful information from physical sensors. Meanwhile, the relatively rich user-contributed postings which contain valuable information about traffic and road conditions was not leveraged to enhance the effectiveness of ITSs. To address such a research gap, this paper presents a relatively new approach of extracting and analyzing useful driving navigation information from the big data archived on online social media, the so-called social sensors. In particular, we advocate a topic model-based computational method to extract relevant semantics (e.g., traffic conditions, road conditions, drivers' conditions, etc.) from the relatively noisy user-contributed postings on online social media. Then, a classification ensemble method is proposed to automatically identify specific traffic related events, and transmit these useful navigation hints back to an ITS for dissemination to other drivers. Evaluation based on real-world user postings from Twitter and Sina Weibo demonstrates that the proposed social sensor analytics method is effective in identifying main traffic events when compared to other baseline methods. Our work opens the door for the development of the next generation of ITSs which can leverage the big data archived on online social media to enhance the richness and the quality of traffic navigation aids provided by ITSs.

I. INTRODUCTION

With the rise of intelligent transportation systems (ITSs), drivers are provided with real-time navigation information (e.g., the shortest driving paths, real-time traffic conditions, traveling time, and other road condition data) to facilitate their driving [1]. These ITSs first collect raw data from physical sensors installed on roads or vehicles, and weather forecasts or event data from various media and transportation control centers. The raw data are then consolidated and analyzed to generate useful driving advices for subsequent dissemination to drivers registered in the ITSs. Previous work mainly examines issues such as the design of efficient ITS frameworks, the design of real-time communication protocols, and the computational methods of extracting useful data from physical sensors to estimate traveling time [2], [3]. However, the relatively rich user-contributed postings which contain valuable information about traffic and road conditions was not



Fig. 1. Traffic Condition Covered by a Tweet



Fig. 2. Traffic Condition Covered by a Weibo Message

leveraged to enhance the effectiveness of ITSs. To address such a research gap, this paper presents a relatively new approach of extracting useful driving navigation information based on the big data archived on online social media, the so-called social sensors, for the generation of useful driving advices in real-time. The notion of social sensors has been explored in various contexts such as earthquake prediction [4], epidemic tracking [5], and traffic monitoring [6]; these sensors have been shown useful for providing vital data for various prediction tasks.

Our proposed approach differs from the previous methodology of leveraging social media data for providing navigation support [6] in that we advocate a topic model-based computational method to extract relevant semantics (e.g., traffic conditions, road conditions, drivers' conditions, etc.) from the relatively noisy user-contributed postings on online social media. In fact, previous research acknowledged the inherent problem of extracting useful and relevant information from noisy social media data [6]. For example, the keyword “heavy traffic” embedded in a tweet may refer to traffic condition on roads, or it may actually point to something totally irrelevant such as “heavy traffic radio” (e.g., the writer of a tweet is listening to the Heavy Traffic Radio station). Our research just aims to address such a problem by using rich semantics captured in topic models to identify and filter noisy postings on online social media such as Tweeter or Sina Weibo. After

semantically rich message filtering, a classification ensemble method is then applied to automatically classify a variety of traffic related events such as road accidents, traffic jams, sudden extreme weather, and so on. Finally, automatically extracted traffic events are consolidated and transferred to other drivers on the road via the proposed ITS. Experiments based on real-world user postings from Twitter and Sina Weibo confirms that the proposed topic modeling-based message filtering method is more accurate than the classical keyword-based filtering method. Moreover, the proposed ensemble-based classification method is more effective than individual classifiers in traffic event classification.

In sum, the main contributions of our research work reported in this paper are as follows: (1) we advocate a social sensor-based framework for extracting vital transportation intelligence from online social media to enhance the effectiveness of ITSs; (2) we develop a topic modeling-based method for filtering noisy social postings so as to extract traffic related messages from online social media; (3) we design an ensemble-based classification method to improve traffic event identification from social sensors; (4) we empirically validate the effectiveness of the proposed framework and computational methods based on real-world user postings extracted from bilingual social media data sets. To the best of our knowledge, this is the first successful research of exploiting a semantically rich topic modeling method for extracting and identifying traffic events from social sensors. The practical implication of our work is that the big data (e.g., huge volume of user-contributed real-time messages) archived on online social media can be tapped into for enhancing the richness and the quality of traffic navigation aids provided by ITSs.

The remainder of the paper is organized as follows. Section II describes the overall framework of the proposed social sensor-based intelligent transportation system. The computational details of the topic modeling-based message filtering method and the classification ensemble-based traffic event identification method are then illustrated in Section III. Section IV describes the evaluation of the proposed framework and computational methods. Finally, we offer concluding remarks and describe future directions of our research work.

II. A SOCIAL SENSOR BASED FRAMEWORK FOR INTELLIGENT TRANSPORTATION

Figure 3 depicts an overview of the proposed framework for social sensor based intelligent transportation system. For the first step, user-contributed messages are extracted from online social media by using standard APIs provided by the source (e.g., the Twitter Streaming API) or dedicated crawler programs. For the second step, a preliminary message filter is applied to extract eligible social postings. For example, a message with a minimal length (e.g., containing at least 10 words) and containing certain named entities (e.g., a location) is selected while a message containing purely numbers, pictures, or URLs are filtered. For named entity recognition (NER), we apply the open source Stanford NER tool [7] for location identification in online postings. For the third

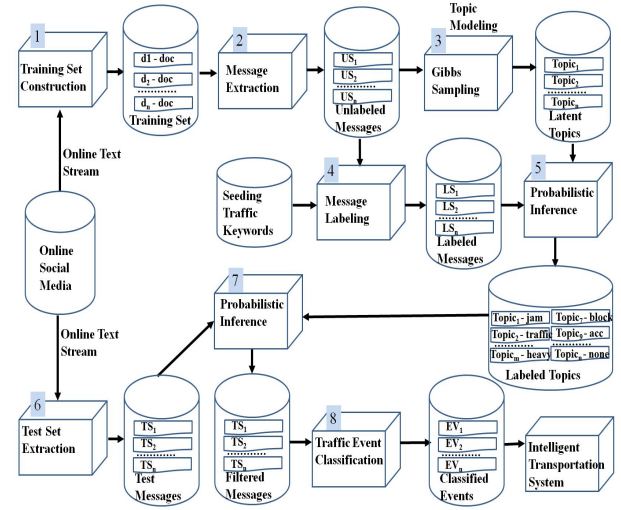


Fig. 3. A Framework of Social Sensor Based ITS

step, the selected unlabeled messages are then fed into a Latent Dirichlet Allocation (LDA) [8] based topic modeling module for unsupervised topic mining. Some of the mined topics contain semantically rich descriptions about various traffic events. To identify traffic relevant events for subsequent event classification, we utilize a seeding set of common traffic event keywords to label a set of user postings in step 4. For the fifth step, a language modeling based probabilistic inference method [9], [10] is applied to infer the label of mined topics based on the labeled messages generated in the forth step. Suppose that a continuous text stream from online social media is extracted by the sixth step. Armed with labeled topics about various traffic events, the language modeling based probabilistic inference method is then applied to identify and filter traffic related messages in the seventh step. Finally, a classification ensemble method is applied to identify specific traffic events. If the classified message comes with a geographic tag (e.g., GPS tag), the exact coordinates of the traffic event location are obtained. If such information is not available in the message, the location name embedded in the message is passed to the Google Map API to extract the relevant coordinates [6]. After obtaining the exact coordinates of the identified traffic events (e.g., road accidents, natural disasters, etc.), this information can then be passed to the ITS for routing to other drivers.

III. THE COMPUTATIONAL METHODS

A. Topic Modeling Based Message Filtering

For the proposed computational framework, Latent Dirichlet Allocation (LDA) [8], a hierarchical Bayesian topic model, is used to learn semantically rich concepts (i.e., topics) about various traffic events from an unlabeled message corpus. In LDA-based topic modeling, each document (message) $d \in D$ of an unlabeled training corpus D is characterized by a multinomial distribution θ , which is sampled from a Dirichlet prior with

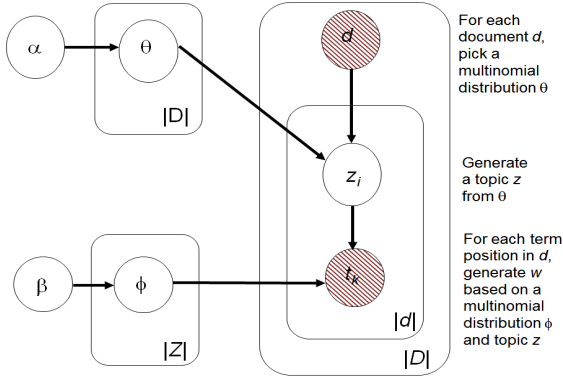


Fig. 4. LDA-based Topic Modeling Process

a hyper-parameter α [8]. Z represents the set of latent topics characterizing D . A latent topic $z_i \in Z$ (i.e., a semantically rich concept about a kind of traffic event) is selected according to the multinomial distribution θ . Given the topic z_i , a term t_k is then generated according to the multinomial distribution ϕ , which is sampled from another Dirichlet prior with a hyper-parameter β . In Figure 4, plate notation is applied to describe the probabilistic generation process of a message corpus. For the plate notation, shaded and unshaded circles indicate observed and latent variables, respectively. Plates (i.e., the rectangles) indicate repeated sampling, with the number of iterations defined by the variable shown at the bottom right corner of each plate.

Our ultimate goal is to infer the conditional probability $Pr(t_k|z_i)$ which represents a latent topic (i.e., a concept). In other words, we want to establish the multinomial distribution ϕ . Nevertheless, directly computing the multinomial distributions ϕ or θ is computationally expensive. Accordingly, various methods, such as variational inference [8], expectation propagation [11], and Gibbs sampling [12] have been applied to approximate ϕ and θ . In our framework, Gibbs sampling [12], a Markov chain Monte Carlo algorithm, is used to estimate the conditional probabilities of the corresponding topic model. For Gibbs sampling, a Markov chain (a model of transitions among successive states) is established by repeatedly drawing a latent topic for each observable term, based on its conditional probability over other variables [11]. Based on previous work [11], [13], the approximations $\bar{\phi}$ and $\bar{\theta}$ of the multinomial distributions ϕ and θ are computed as follows:

$$\bar{\phi} = \frac{C_{mn}^{VZ} + \beta}{\sum_{m' \in V} C_{m'n}^{VZ} + |V|\beta} \quad (1)$$

$$\bar{\theta} = \frac{C_{np}^{ZD} + \alpha}{\sum_{n' \in Z} C_{n'p}^{ZD} + |Z|\alpha} \quad (2)$$

where C_{mn}^{VZ} is a count matrix that captures the number of times that term $t_k = m$ is assigned to latent topic $z_i = n$, excluding the current term position. The notations m and n represent the indices in the count matrix C_{mn}^{VZ} . The set of vocabulary V is

used to compose the collection D , and the set Z represents the set of latent topics characterizing the collection D . For our proposed topic model, the vocabulary V contains all the terms appearing in the corpus. The count matrix C_{np}^{ZD} records the number of times that the latent topic $z_i = n$ is assigned to a document $d_j = p$, excluding the current document under consideration. The terms n and p are the indices of the count matrix C_{np}^{ZD} . In Eq. 1, the term $\frac{C_{mn}^{VZ} + \beta}{\sum_{m' \in V} C_{m'n}^{VZ} + |V|\beta}$ is used to approximate the probability of the term t_k given the latent topic z_i , whereas $\frac{C_{np}^{ZD} + \alpha}{\sum_{n' \in Z} C_{n'p}^{ZD} + |Z|\alpha}$ is applied to approximate the probability of z_i given the document d_j in Eq. 2.

With reference to Figure 4, the probability of z_i given the observed variables t_k and d_j is computed as follows.

$$\begin{aligned} Pr(z_i|t_k, d_j) &\approx \\ Pr(t_k = m|z_i = n) \cdot Pr(z_i = n|d_j = p) &\approx \end{aligned} \quad (3)$$

$$\frac{C_{mn}^{VZ} + \beta}{\sum_{m' \in V} C_{m'n}^{VZ} + |V|\beta} \cdot \frac{C_{np}^{ZD} + \alpha}{\sum_{n' \in Z} C_{n'p}^{ZD} + |Z|\alpha}$$

During model learning, the count matrix C_{mn}^{VZ} which captures $V \times Z$ (term by topic) and the count matrix C_{np}^{ZD} that holds $Z \times D$ (topic by document) are maintained and updated in each Gibbs iteration. Intuitively, if a topic z_i has been assigned to many documents across the whole corpus D , the probability of the topic z_i being assigned to an arbitrary document d_j increases. Similarly, if a topic z_i has occurred many times in a document d_j , the probability of an arbitrary term being assigned to the topic z_i in d_j increases. Essentially, iterations of the Gibbs sampling process just try to capture the aforementioned intuition. As a whole, the likelihood of a term t_k being assigned to a topic z_i not only depends on how likely the term t_k is associated with the topic z_i but also the likelihood of the topic z_i occurs in the entire corpus D . Each Gibbs sample consists of topic assignments to all the terms in the corpus. The samples generated before the “burn-in” period (e.g., the first 500 iterations of the Gibbs sampling algorithm) are usually ignored because of the low quality of these estimations [14]. After the “burn-in” period, successive Gibbs samples converge to produce a good estimation of the posterior distributions over topic assignments. In other words, the Gibbs samples produced after the “burn-in” period should be saved to compute $\bar{\theta}$ and $\bar{\phi}$. We extend the open source Core LingPipe API (<http://alias-i.com/lingpipe/web/download.html>) to develop the Gibbs sampling algorithm for the implementation of the proposed topic model.

B. Probabilistic Language Modeling Based Inference

We propose a probabilistic language model (LM) based method to estimate the generation probabilities of various labeled message based on a mined unlabeled topic. A language model is a probabilistic function that assigns a probability to a string t drawn from some vocabulary T . Language modeling has been applied to estimate the relevance of a document d with respect to a query q in terms of the likelihood of in the field of IR [10]. The basic unigram language model is defined by [9], [10]:

$$Score_{LM}(q, d) = P(q|d) = \prod_{t_i \in q} P(t_i|M_d) \quad (4)$$

where $Score_{LM}(q, d)$ represents the document score of d with respect to the query q , whereas M_d is the probabilistic language model of the document d . To replace the zero probability of an un-seen query term by a small value rather than zero, Jelinek-Mercer smoothing [9] was proposed. A smoothed document language model is defined by:

$$P(t_i|M_d) = (1 - \lambda)P_{ML}(t_i|M_d) + \lambda P_{ML}(t_i|M_D) \quad (5)$$

$$P_{ML}(t_i|M_D) = \frac{tf(t_i, D)}{|D|} \quad (6)$$

where $P_{ML}(t_i|M_D)$ is the maximum likelihood estimation of a collection language model. λ is the Jelinek-Mercer smoothing parameter, which may take values in the range of [0.1, 0.7] [15], [16]. $tf(t_i, D)$ represents the occurrence frequency of t_i in the entire document collection D (i.e., all the consumer reviews).

Finally, for each mined topic, we apply Eq. 5 to Eq. 7 to compute the generation probabilities of z_i with respect to each label message LS_i . If a topic with generation probability greater than a minimal threshold for a label i , and its probability is higher than other generation probabilities for other label j , it will be inferred with label i .

$$Score_{LM}(LS_i, z_i) = P(LS_i|z_i) = \prod_{t_i \in LS_i} P(t_i|M_{z_i}) \quad (7)$$

C. Classification Ensemble for Event Classification

One of the base classifiers we used is support vector machines (SVM) which have been widely used in text classification tasks. Its performance has been shown better than that of probabilistic classifiers such as Naive Bayes [17]. For the two-class classification problem (e.g., traffic jam or not), the basic principle is to find a hyperplane represented by the weight vector (a normal vector perpendicular to the hyperplane), which not only separates the message vectors into two classes, but also guarantees the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem:

$$\vec{\omega} = \sum \alpha_i c_i \vec{d}_i, \forall \alpha_i \geq 0 \quad (8)$$

$$b = c_i - \vec{\omega}_i^T \vec{d}_i, \forall \vec{d}_i : \alpha_i \neq 0 \quad (9)$$

where each $\vec{d}_i$ represents a labeled training example (i.e., labeled message), and $c_i \in \{-1, +1\}$ denotes a class of such a binary classification problem. The term α_i is the Lagrange multiplier, and each nonzero α_i indicates that the corresponding $\vec{d}_i$ is a support vector; b is the intercept of the binary classifier. Based on the parameters such as ω_i , α_i , and b with appropriate values established through a training

set, the large margin classification function $f(\vec{d}_i)$ can then be defined, where the signum function (sgn) is applied to decide the class label of a given test instance $\vec{d}$ (i.e., an unlabelled user-contributed message). For example, a positive margin value indicates a traffic jam, and a negative margin value indicates a normal traffic condition. We applied the Spark Machine Learning (MLib) library ¹ to implement the SVM classifier with a linear kernel function and adopted all the default parameter values.

$$f(\vec{d}) = \text{sgn} \left(\sum_i \alpha_i c_i \vec{d}_i^T \vec{d} + b \right) \quad (10)$$

The other two base classifiers we used include a Naive Bayes classifier and a K-Nearest Neighbor (KNN) classifier [18]. We then develop our ensemble classifier based on a weighted voting scheme. The final classification score is computed based on the following weighted classification function of the three base classifiers:

$$f(\text{ensem}) = \sum_i \omega_i \times f(\text{class}_i) \quad (11)$$

where $f(\text{ensem})$ is the ensemble classification function, and ω_i and $f(\text{class}_i)$ are the weight and the base classification function of the classifier i , respectively. For a binary classification task, the ensemble's decision is based on the majority class label assigned by the individual classifiers.

IV. EXPERIMENTS AND RESULTS

To evaluate the proposed methods for filtering noisy messages and identifying traffic events from social sensors, we performed two series of experiments based on two data sets. We retrieved 2 million user postings from Twitter by using a Twitter scraper. These tweets were posted between January 2014 and December 2014. In addition, we also downloaded 1.8 million user postings pertaining to the same period from Sina Weibo by using our crawler program. We then select a subset of messages from each data set for evaluation data labeling. For the Twitter data set, we successfully labeled three kinds of events such as traffic jams (400 messages), poor road conditions (400 messages), and traffic restrictions (200 messages). In addition, there are other 1,000 messages which are confirmed not related to any traffic events at all. Labeling were conducted by two researchers engaging in research on intelligent transportation systems for several years. Each label in the evaluation data set was confirmed only when both annotators agreed on the particular label assignment. Similarly, these two annotators also labeled the same kinds of messages for the Sina Weibo data set. At the end, there are 350 messages about traffic jam, 350 messages about poor road conditions, 300 messages about traffic restrictions, and 1,000 messages not related to any traffic events.

For the message filtering experiments, the experimental system is the proposed topic modeling-based method and the

¹<http://spark.apache.org/>

TABLE I
RESULTS OF THE MESSAGE FILTERING EXPERIMENTS

Dataset	Experimental		Baseline		F-score
	Recall	Precision	Recall	Precision	Improvement
Twitter	0.825	0.789	0.655	0.824	+10.56%
Weibo	0.645	0.746	0.525	0.789	+9.68%
Average	0.735	0.768	0.590	0.807	+10.12%

baseline system is the vector space model (i.e., bag of words approach) [19]. The objective of this experiment is to test if the topic modeling-based method can filter traffic related message or not. There are only two classes for the filtering tasks namely, traffic related messages or noisy (not traffic related) messages. At the beginning, the whole message corpus (e.g., the Twitter corpus) was fed to the proposed LDA model for topic learning. Then, the probabilistic language model (Eq. 5 to Eq. 7) was applied to infer the labels of these topics based on a set of labeled messages constructed by using a small seeding set of traffic related keywords. The vector space baseline model utilized the same seeding set of traffic related keywords for message filtering. A 5-fold cross validation was applied to the filtering experiments. The common performance measures such as recall, precision, and F-score were applied to evaluate the performance of both systems. The performance of these systems is reported in Table I. The values shown in the table represent the average performance scores generated from the 5-fold of evaluation. Similarly, we applied more or less the same experimental procedure to evaluate the systems with respect to the Weibo data set. However, a Chinese word segmentation system named ICTCLAS² developed by the Chinese Academy of Sciences was applied to identify meaningful word boundaries before we invoked the respective systems.

According to the results shown in Table I, the experimental system (i.e., topic modeling-based message filter) consistently performs better than the baseline vector space model in terms of F-score across different data sets. The reason is that the proposed topic modeling method can extract semantically rich topics (concepts) characterizing different kinds of traffic events in a semi-supervised manner (a manually constructed seeding set of traffic terms is needed). Armed with semantically rich traffic related topics learned from a large corpus of unlabeled messages, the proposed method can more effectively filter out noisy messages from the test set. Since the vector space model utilized manually identified seeding set of traffic keywords for message filtering, the precision of the filtering process is generally high. However, manually identified traffic terms are limited and are likely to be incomplete. As a result, the recall of the vector space model is very poor. The experimental results confirm that the proposed topic modeling method for traffic related message filtering is effective, and it achieves an average performance improvement of +10.12%.

For the second series of experiments, we aimed to evalu-

TABLE II
RESULTS OF TRAFFIC EVENT IDENTIFICATION FROM SOCIAL SENSORS

Method	Twitter			Weibo		
	Recall	Precision	F-score	Recall	Precision	F-score
Traffic Jams						
Ensemble	0.788	0.768	0.778	0.743	0.722	0.732
SVM	0.738	0.747	0.742	0.757	0.716	0.736
KNN	0.763	0.718	0.739	0.729	0.708	0.718
NB	0.688	0.724	0.705	0.686	0.696	0.691
Poor Road Conditions						
Ensemble	0.775	0.765	0.770	0.757	0.726	0.741
SVM	0.725	0.744	0.734	0.743	0.712	0.727
KNN	0.750	0.714	0.732	0.714	0.704	0.709
NB	0.675	0.720	0.697	0.671	0.691	0.681
Traffic Restrictions						
Ensemble	0.763	0.753	0.758	0.729	0.718	0.723
SVM	0.713	0.750	0.731	0.729	0.708	0.718
KNN	0.738	0.728	0.733	0.714	0.725	0.719
NB	0.663	0.726	0.693	0.686	0.676	0.681

ate the effectiveness of the proposed classification ensemble method for traffic event identification from social sensors. In each test run, we examined the classification performance for a single kind of traffic event, and so it is a binary classification task. The baseline systems include the SVM [17], KNN [18], and Naive Bayes classifiers. For all classification methods, we employed the word unigram features. For the proposed weighted voting scheme, we set equal weight for each base classifier. Similarly, a 5-fold cross validation method was employed in this series of experiments. The performance of various methods across different data sets is shown in Table II. Our experimental results reveal that the proposed ensemble method outperforms all baseline systems in all kinds of traffic event identifications for the Twitter data set. Moreover, the proposed method outperforms most baseline systems for different traffic event identification for the Weibo data set. The main reason is that the ensemble method can leverage the merits of individual classifiers based on the weighted voting scheme. Consequently, the classification results are more accurate. The reason why the ensemble method does not significantly outperform the SVM method for the Weibo data set may be due to the challenge in accurate Chinese word segmentation. If the Chinese words were not properly segmented, the corresponding topical features might contain a lot of noise, which led to poor classification results. As a whole, we demonstrate the advantage of the proposed classification ensemble method in both English and Chinese traffic event identification in these experiments.

V. CONCLUSIONS AND FUTURE WORK

Though relatively rich traffic information may be extracted from the so-called social sensors to enhance ITSs, the messages posted to online social media are generally noisy. Moreover, it is a very challenging problem of effectively identifying various traffic events from free texts. This paper examines a new method to address the existing research gaps, and presents a topic model-based method for message filtering and an ensemble classification method for traffic event identification

²<http://ictclas.org/>

from social sensors. Based on two real-world social media data sets, our experimental results confirm that the proposed semantically rich topic modeling method for message filtering is effective, and it outperform the classical keyword-based method by 10% in terms of F-score. The proposed classification ensemble method for traffic event identification also outperforms classical machine learning classifiers in most traffic event identification tasks across different data sets. Future work involves the use of topical features extracted from the proposed topic model to further bootstrap the classification performance of the proposed ensemble method in traffic event identification.

ACKNOWLEDGMENTS

This work is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.: CityU 11502115), the Shenzhen Municipal Science and Technology R&D Funding - Basic Research Program (Project No.: JCYJ20160229165300897), and Natural Science Foundation China, NSFC (Project No.: 71671155).

REFERENCES

- [1] X. Cheng, L. Yang, and X. Shen, "D2D for intelligent transportation systems: A feasibility study," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1784–1793, 2015. [Online]. Available: <http://dx.doi.org/10.1109/TITS.2014.2377074>
- [2] P. Kachroo and S. Sastry, "Traffic assignment using a density-based travel-time function for intelligent transportation systems," *IEEE Trans. Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1438–1447, 2016. [Online]. Available: <http://dx.doi.org/10.1109/TITS.2015.2503348>
- [3] —, "Travel time dynamics for intelligent transportation systems: Theory and applications," *IEEE Trans. Intelligent Transportation Systems*, vol. 17, no. 2, pp. 385–394, 2016. [Online]. Available: <http://dx.doi.org/10.1109/TITS.2015.2472556>
- [4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds. ACM, 2010, pp. 851–860. [Online]. Available: <http://doi.acm.org/10.1145/1772690.1772777>
- [5] E. Diaz-Aviles and A. Stewart, "Tracking twitter for epidemic intelligence: case study: EHEC/HUS outbreak in germany 2011," in *Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012*. ACM, 2012, pp. 82–85. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2380718>
- [6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Real-time event extraction for driving information from social sensors," in *Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*. IEEE, 2012, pp. 221–226.
- [7] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 363–370. [Online]. Available: <http://www.aclweb.org/anthology/P/P05/P05-1045>
- [8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: <http://www.jmlr.org/papers/v3/blei03a.html>
- [9] X. Liu and B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 186–193.
- [10] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, pp. 275–281.
- [11] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, Eds. Seattle, Washington: ACM, Aug. 22–25 2004, pp. 306–315. [Online]. Available: <http://doi.acm.org/10.1145/1014052.1014087>
- [12] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian relation of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.
- [13] M. Rosen-Zvi, C. Chemudugunta, T. L. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems*, vol. 28, no. 1, pp. Article–4, 2010. [Online]. Available: <http://doi.acm.org/10.1145/1658377.1658381>
- [14] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 178–185. [Online]. Available: <http://doi.acm.org/10.1145/1148170.1148204>
- [15] J.-Y. Nie, G. Cao, and J. Bai, "Inferential language models for information retrieval," *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 4, pp. 296–322, Dec. 2006.
- [16] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems*, vol. 22, pp. 179–214, 2004.
- [17] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, ser. Lecture Notes in Computer Science, vol. 1398. Springer, 1998, pp. 137–142. [Online]. Available: <http://dx.doi.org/10.1007/BFb0026683>
- [18] Y. Gao and F. Gao, "Edited adaboost by weighted kNN," *Neurocomputing*, vol. 73, no. 16-18, pp. 3079–3088, 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2010.06.024>
- [19] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York, New York: McGraw-Hill, 1983.