



Zürich, August 3, 2016

A Data Structure for Group-Based Temporal Probabilistic Operations

A tuple (F, T, λ, p) in a temporal probabilistic database (TPDB) states that at each time point in T , fact f is *true* with probability p and *false* with probability $1 - p$. Attribute λ is a boolean formula consisting of tuple identifiers and logical symbols, used to connect these variables. The role of lineage is to indicate "how and based on which input tuples has an output tuple been derived".

r (productsWanted)						s (productsInStock)					
k	P	B	T	λ	p	k	P	B	T	λ	p
r_1	m	b	[2,10)	r_1	0.3	s_1	m	b	[1,4)	s_1	0.6
r_2	c	b	[4,7)	r_2	0.2	s_2	m	b	[6,8)	s_2	0.3
r_3	d	b	[1,3)	r_3	0.4	s_3	c	b	[4,5)	s_3	0.7
						s_4	c	b	[7,9)	s_4	0.1

Figure 1: The Supermarket Application Scenario

As an example of a TPDB, consider a supermarket that, given the shopping data of its clients and data related to its inventory, for each time point and store, predicts the products that clients want to buy and the merchandise in stock. In the above figure, relation r contains the products (P) from brand (B) that customers buy during a month, based on their shopping habits. Relation s contains the products (P) from brand (B) that are planned to be in stock. For example, tuple r_1 from relation r states that, at each day, from the 2nd until the 10th of the month, "product m from brand b is bought" with probability 0.3. Tuple s_1 from relation s states that, at each day from the 1st until the 4th, "product m from brand b is available" with probability 0.4. The lineage expressions of all the tuples in r and s are equal to their identifiers since they are all base tuples.

Assume a set of temporal probabilistic (TP) tuples belonging to either one or two relations and a group-based TP operation (projection, aggregation, set-operations). The tuples valid over each subinterval T will determine the output tuple over T , its lineage expression and its

probability. The goal of this project is to implement a data structure (e.g. list, tree) that could store the valid tuples over each output interval and use it to perform all group-based temporal probabilistic operations. This structure needs to maintain its contents in a sorted form so that the tuples valid at a time-point can be fetched, inserted and deleted efficiently.

Tasks

1. Study and evaluation of the data structures used in the related work [1, 3, 2] .
2. Implementation of a data structure for group-based TP operations in PostgreSQL.
3. Implementation of the corresponding algorithm that can produce the result of all group-based TP operations in PostgreSQL using the chosen data structure.
4. Experimental evaluation and comparison with existing approaches.
5. Written thesis (approximately 50 pages)
6. 25-minute Presentation of the results in a group meeting.

References

- [1] Michael H. Böhlen, Johann Gamper, and Christian S. Jensen. Multi-dimensional aggregation for temporal data. In *EDBT*, 2006.
- [2] Martin Kaufmann, Panagiotis Vagenas, Peter M. Fischer, Donald Kossmann, and Franz Färber. Comprehensive and interactive temporal query processing with sap hana. *PVLDB*, 2013.
- [3] Markus Pilman, Martin Kaufmann, Florian Köhl, Donald Kossmann, and Damien Profeta. Partime: Parallel temporal aggregation. In *SIGMOD*, 2016.

Supervisor: Katerina Papaioannou

Start date: As soon as possible

End date: 6 months after the starting date

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen