University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Student: Mirko Richter

Group-Based Temporal Probabilistic Operations [BSc Thesis]

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)

k	P	B	T	λ	p
r_1	m	b	[2,10)	r_1	0.3
r_2	c	b	[4,7)	r_2	0.2
r_3	d	b	[1,3)	r_3	0.4

s (productsInStock)

k	P	B	T	λ	p
s_1	m	b	[1,4)	s_1	0.6
s_2	m	b	[6,8)	s_2	0.3
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 2^{nd} until the 10^{th} 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 1^{st} until the 4^{th}, "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 group-based TP operation (projection, aggregation, set-operations) applied on one or both of the two given relations. 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 sweepline based algorithm that would update the set of valid tuples over each output interval and use the corresponding lineage expressions to produce the output one lineage expression.

Tasks

1. Implementation of a sweeping algorithm that is appropriate for computing all group-based TP operations.
2. Study and implementation of approaches used in the related work [1, 2, 3] for the computation of group-based operations in temporal databases.
3. Experimental evaluation and comparison with existing approaches.
4. Written thesis (approximately 50 pages)
5. 25-minute Presentation of the results in a group meeting.

Optional: Incorporation of the sweeping algorithm in PostgreSQL.

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: 30-01-2017
End date: 30-07-2017
Presentation date: 04-07-2017

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen