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

Andreas Albrecht

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

Zürich, April 4, 2013

BSc Thesis
Topic: Learning Value Evolution on Real-world Temporal Data

Many data sets contain temporal records over a long period of time; each record is associated with a time stamp and describes some aspects of a real- world entity at that particular time (e.g., author information in DBLP[1]). In such cases, we often wish to know the history of an entity and so be able to enable interesting longitudinal data analysis. For example, DBLP lists research papers over many decades; DBLP users may wish to find authors by name, find the publication history and affiliation of an author, find her research topics over time, and so on.

A major challenges for enabling such search and exploration is to identify records that describe the same real-world entity over a long period of time; only with such an integrated view, we will be able to trace the history of that entity and collect statistics over time. However, linking temporal records is by no means easy. First, entities can evolve over time, thus, records that describe the same real-world entity at different times can contain different values; for example, a researcher can move from one affiliation to another. Second, records that describe different entities at different times can share common values; for example, having two persons with highly similar names in the same university over the past 30 years is more likely than at the same time. We illustrate the challenges by the following example.

EXAMPLE 1. Consider records that describe paper authors in Table 1; each record is derived from a publication record at DBLP. These records describe 3 real-world persons: r_1 describes E_1: *Xin Dong*, who was at *R. Polytechnic* in 1991; $r_2 - r_6$ describe E_2: *Xin Luna Dong*, who moved from *Univ of Washington* to *AT&T Labs*; $r_7 - r_{12}$ describe E_3: *Dong Xin*, who moved from *Univ of Illinois* to *Microsoft Research*.

The key to correctly link records in Example 1 is to understand how values of real-world entities

[1] http://www.dblp.org/

Table 1: Records from DBLP.

ID	name	affiliation	co-authors	year
r_1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r_2	Xin Dong	Univ of Washington	Halevy, Tatarinov	2004
r_3	Xin Dong	Univ of Washington	Halevy	2005
r_4	Xin Luna Dong	Univ of Washington	Halevy, Yu	2007
r_5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r_6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r_7	Dong Xin	Univ of Illinois	Han, Wah	2004
r_8	Dong Xin	Univ of Illinois	Wah	2007
r_9	Dong Xin	Microsoft Research	Wu, Han	2008
r_{10}	Dong Xin	Univ of Illinois	Ling, He	2009
r_{11}	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r_{12}	Dong Xin	Microsoft Research	Ganti	2010

evolve over time. For instance, we should be able to answer the following questions. (1) What is the probability that *Xin Luna Dong* moved to another affiliation after 5 years in *Univ. of Washington*? (2) What is the probability that two records, *e.g.* r_1 and r_2 with the same name *Xin Dong* and a time gap of 13 years, refer to the same author? Answers to the above questions are captured by decay patterns of real-world entities [3].

Within this project, the student should be able to compute affiliation decay patterns on sampled publication data from DBLP, where we know the ground truth of underlying authors. Specifically, the patterns covers two aspects: (1) the probability of an author changing affiliations within a given time distance; (2) the probability of different authors sharing the same affiliation within a given time distance.

Tasks

1. Understand time decay model and string similarity computation approaches in the literature [1, 2, 3].
2. Develop the proposed algorithms in [3] to learn time decay patterns, including disagreement decay and agreement decay.
3. Implement the decay-learning algorithms on affiliation attribute of 2000 sampled paper records from DBLP.
4. Write thesis (approximately 50 pages).
5. Present the results at group meeting (maximal 25 min).

Supervisor:	Pei Li (peili@ifi.uzh.ch)
Start date:	03.04.2013
End date:	31.07.2013
Duration:	4 months

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen

References

[1] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *VLDB J.*, 18(1):255–276, 2009.

[2] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[3] Pei Li, Xin Luna Dong, Andrea Maurino, and Divesh Srivastava. Linking temporal records. *PVLDB*, 4(11):956–967, 2011.