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

Gionata Genazzi

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

Zürich, January 27, 2015

BSc Thesis
Topic: Temporal Filtering to Improve Temporal Duplicate Detection

Given a data set $\mathbf{R}$, duplicate detection is the activity of identifying all record pairs referring to the same real-world entity in $\mathbf{R}$. A quantitative way of duplicate detection is to use a similarity function, and perform a *similarity join* to find all pairs of records whose similarities are above a given threshold. An algorithmic challenge is how to perform the similarity join in an *efficient* and *scalable* way.

Xiao et al. [3] proposed a filtering technique, *positional filtering*, to efficiently find pairs of records such that their Jaccard similarities are above a given threshold. Specifically, the proposed algorithm exploits the ordering of tokens in a record and leads to upper bound estimates of Jaccard similarity scores.

Consider two records x and y. Jarccard similarity of x and y, denoted as $J(x, y)$, is defined as $\frac{|x \cap y|}{|x \cup y|}$. Let $O(x, y)$ be the overlap similarity $|x \cap y|$. Given a similarity threshold θ, positional filtering algorithm estimates an upper bound of Jaccard similarity by the following property:

$$J(x, y) \geq \theta \Longleftrightarrow O(x, y) \geq \frac{\theta}{1+\theta}(|x| + |y|) \tag{1}$$

We show the effect of positional filtering in the following example.

EXAMPLE 1. Consider two authors described by their co-authors as:

$$r_x = \text{“}Agrawal, Cormode, Gebaly, Golab, Korn\text{”}$$

$$r_y = \text{“}Cormode, Gebaly, Golab, Korn, Valluri\text{”}$$

They can be transformed into the following two records:

$$x = [A, B, C, D, E]$$

$$y = [B, C, D, E, F]$$

The Jaccard similarity of x and y is $\frac{4}{6} = .67$. Consider similarity threshold $\theta = .8$. Positional filtering prunes record pair (x, y) from similarity join as follows. Let all tokens be sorted by a global order (e.g., lexicographic order), and the first two tokens in each record be their prefixes. Records x and y share token B in their prefixes. Thus, an estimate of the maximum possible overlap can be obtained as the sum of overlap amount in prefixes and the minimum number of unseen tokens in x and y, i.e., $1 + min(3, 4) = 4$. Given the Jaccard similarity property, thesshold $\theta = .8$ means that overlap $O(x, y)$ between x and y should be at least 5. Thus, record pair (x, y) can be pruned.

Study on temporal datasets brings new challenges to duplicate detection. First, 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. Recent work [2, 1] proposed *time decay models* to capture the effect of time elapse on entity value evolution. Applying decay when computing similarity between temporal records improves accuracy over traditional duplicate detection techniques.

EXAMPLE 2 Consider two authors as in **Example 1**, each associated with a timestamp.

$$r_x = \text{“}Agrawal, Cormode, Gebaly, Golab, Korn\text{”}, 2004$$

$$r_y = \text{“}Cormode, Gebaly, Golab, Korn, Valluri\text{”}, 2009$$

Consider Jaccard similarity with threshold $\theta = .8$. Temporal decay refines Jaccard similarity as follows. It assigns value weights to tokens of two records. Value agreement between two records with a large time distance is less rewarded by a weight less than 1; value disagreement between two records with a large time distance is less penalized by a weight less than 1. Suppose the weight $w_{agr}(\Delta t = 5)$ of value agreement over 5 years is $.9$, and the weight $w_{dis}(\Delta t)$ of value disagreement over 5 years is $.2$. Thus *temporal Jaccard* similarity between records x and y is computed as

$$\frac{.9 + .9 + .9 + .9}{.9 + .9 + .9 + .9 + .2 + .2} = .9$$

Given the similarity threshold $\theta = .8$, record pair (x, y) should not be pruned by positional filtering algorithm.

In this project, the student is to refine the positional filtering algorithm for *temporal Jaccard* similarity, i.e., the refined positional filtering algorithm is supposed to estimate a correct upper bound of *temporal Jaccard* similarity. The results will be evaluated on a set of DBLP data.

Tasks

1. Understand Jaccard similarity, positional filtering in similarity join [3] and temporal decay model for temporal similarity [2, 1].
2. Implement decay functions using sampled DBLP data set.
3. Given a similarity threshold θ, refine positional filtering algorithm such that each record pair pruned by the algorithm has a temporal Jaccard similarity less than θ.
4. Implement the refined algorithm on a set of DBLP data and verify the results.
5. Write thesis (approximately 50 pages).
6. Present the results at group meeting (maximal 25 min).

Supervisor:	Pei Li (peili@ifi.uzh.ch)
Start date:	26.01.2015
End date:	26.07.2015
Duration:	6 months

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen

References

[1] Yueh-Hsuan Chiang, AnHai Doan, and Jeffrey F. Naughton. Modeling entity evolution for temporal record matching. In *SIGMOD '14*, pages 1175–1186.

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

[3] Chuan Xiao, Wei Wang, Xuemin Lin, and Jeffrey Xu Yu. Efficient similarity joins for near duplicate detection. In *Proceedings of the 17th International Conference on World Wide Web*, pages 131–140, 2008.