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

Tobias Ammann

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

Zürich, March 21, 2014

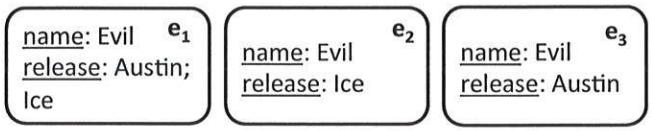
Facharbeit
Topic: Applying Meta-Blocking to Improve Efficiency in Entity Resolution

High efficiency is a mandatory request when handling vast volume of datasets. For example, a simple pairwise comparison in entity resolution takes time $O(n^2)$, where n is the number of entity mentions. To reduce the number of comparisons in entity resolution, we often apply blocking techniques in the pre-processing phase of entity resolution.

Most existing blocking techniques are designed to relational data and face several challenges to data on the Web. Existing blocking strategies usually assume that mentions of the same real-world entity are consistent in one or more attribute, and thus generate blocking keys to cluster entity mentions by specific combinations of these attributes. However, data on the Web are often from multiple sources that may conflict with each other, therefore it is difficult to find a perfect combination of attributes as blocking key. In addition, most existing blocking techniques heavily depend on the existence of a priori known schema. This assumption is broken by poorly structured data on the Web. Recent work [1] solves the problem by *Meta-Blocking* method, which intervenes between the creation and the processing of blocks, transforming an initial set of blocks into a new one with substantially fewer comparisons and equally high effectiveness.

In essence, Meta-Blocking aims at extracting the most similar pairs of entities by leveraging the information that is encapsulated in the block-to-entity relationships. To this end, it first builds

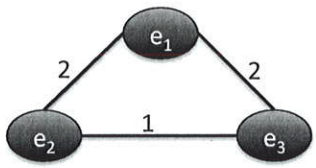
an abstract graph representation of the original set of blocks, with the nodes corresponding to entity mentions and the edges connecting the co-occurring ones. During the creation of this structure all redundant comparisons are discarded, while the false positive paris can be removed by pruning of the edges with the lowest weight. We illustrate the procedure using one of edge weighting schemes and graph pruning algorithms proposed in [1].



(a) artist entity mentions

token	entity mentions
Evil	e_1, e_2, e_3
Austin	e_1, e_3
Ice	e_1, e_2

(b) inverted index



(c) meta-blocking graph

Example 1: Consider the three artist mentions in Figure (a), each described by artist name and a set of releases. To facilitate the process of Meta-Blocking, an inverted list is created (Figure (b)), where each entry of the list records a distinct token in the data and the set of entity mentions sharing the token. With a scan of the list, a Meta-Blocking graph is generated (Figure (c)), where each node represents an entity mention, and an edge connects two nodes that share at least one token. The weight on an edge is the number of tokens two nodes share. To prune the graph, we remove edges with weight below the average weight (1.6), which means we remove the edge between node e_2 and e_3. The results of Meta-Blocking in the dataset are two entity mention pairs (e_1, e_2) and (e_1, e_3), which are considered as candidate matching pairs.

Within this project, the student should be able to understand the Meta-Blocking algorithm, and implement it on a real-world dataset of 800 K artist mentions. Each artist mention is described by name, type, gender, begin year, end year, area and alias. The dataset contains duplicate artist mentions, where the ground truth is known.

Tasks

1. Understand the Meta-Blocking method proposed in [1].
2. Implement the Meta-Blocking method with a selected edge weight schema, a graph pruning algorithm as well as a pruning criteria on the aforementioned dataset.

3. Measure accuracy of the method on the dataset by *precision, recall* and *F-measure.*
4. Measure scalability of the method by runtime on subsets of the data with creasing sizes.
5. Present the results in a written report.

Supervisor:	Pei Li (peili@ifi.uzh.ch)
Start date:	20.03.2014
End date:	12.05.2013
Duration:	2 months

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen

References

[1] George Papadakis, Georgia Koutrika, Themis Palpanas, and Wolfgang Nejdl. Meta-blocking: Taking entity resolution to the next level. *IEEE Transactions on Knowledge and Data Engineering*, 99.