A generic method for entity resolution in databases

With a case study on the cleaning of scientific references in bibliographic databases

Emiel Caron & Wen Xin Lin

School of Economics & Management, Tilburg University, The Netherlands

{e.a.m.caron|w.x.lin}@tilburguniversity.edu

Introduction

Many databases contain ambiguous and unstructured data which makes the information it contains difficult to use. In order for these databases to be a reliable point of reference, the data needs to be cleaned. *Entity resolution* focuses on disambiguating records that refer to the same entity. In this paper we study different applications of entity resolution and propose a generic method for cleaning large bibliographic database, based on [1, 2]. We carry out a Python implementation of this disambiguation method on a table with scientific references in the database PatStat of the European Patent Office [3]. The table holds information on citations to scientific references. While the table could be an important source for research, it typically contains many name variants of the same publications.

Method

Two entities are called *similar* if their corresponding feature vectors f are similar. To measure the similarity of a pair $(f^i(l), f^j(l))$, we define

$$\sigma_l(f^i(l), f^j(l)), \tag{1}$$

where $f^i = (f^i(1), ..., f^i(M))$ denotes the feature vector of an entity r_i. Since we want to obtain a similarity score as a single number, we define a weight vector $\boldsymbol{w} = (w_1, ..., w_M)$ such that

$$s_{ij} = \sum_{k=1}^{M} w_k \cdot \sigma_k(f^i(k), f^j(k)), \tag{2}$$

where $w_k \geq 0, i \neq j$. Similar entities are grouped into sets $\Sigma_1, ..., \Sigma_Q$ as follows. We define a threshold δ such that

$r_j \in \Sigma_p$ implies $\exists r_i \in \Sigma_p$ such that $s_{ij} \geq \delta, i \neq j$

unless $|\Sigma_p| = 1$. The sets are mutually exclusive and are chosen to be maximal:

Let $r_i \in \Sigma_p$. If $\exists r_j$ such that $s_{ij} \geq \delta$, then $r_j \in \Sigma_p$

Research objective

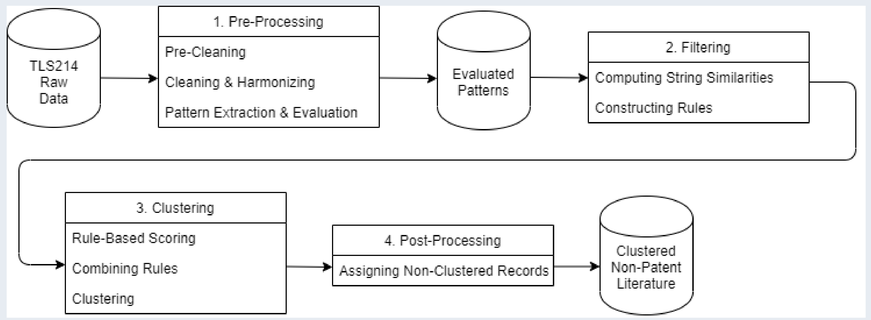

Figure 1. Overview of the disambiguation method

The main goal of our research is to find a generic method for disambiguating large databases and optimize the performance of this method.

Thus our resulting sets

$$\cup_{p=1}^{Q} \Sigma_p = \{r_1, ...r_N\}$$

Evaluation of the performance of our sets is done using a precision and recall analysis. Finally, we maximize the F1-score of the sets by choosing $\boldsymbol{w} = (w_1, ..., w_M)$ and δ in such a way that

$$\boldsymbol{w}^*, \delta^* = \arg\max_{\boldsymbol{w}, \delta} L(\boldsymbol{w}, \delta), \tag{3}$$

where $L(\boldsymbol{w}, \delta)$ is the average F1-score of the sets. Optimizing the parameters is done using a simulated annealing algorithm.

Results

The results of our method, depicted in Figure 2, are promising in terms of precision-recall analysis.

Category	Count	Precision		Recall		F1-score	
		Average	Median	Average	Median	Average	Median
Large (>100)	31	0.9979	1	0.9893	1	0.9933	1
Medium (11-100)	43	1	1	0.9490	1	0.9563	1
Small (<11)	41	1	1	0.5936	1	0.6114	1
Total	115	0.9994	1	0.8331	1	0.8433	1

Figure 2. Statistics of optimized clusters

The top-left graph in figure 3 shows the increase of the F1-score against time. An important conclusion to draw from this is that the simulated annealing algorithm indeed is able to improve the F1-score and thus provide better clusters. The precision and recall results are significantly improved as precision is maximal for 98.3% of the clusters and average recall is 0.83.

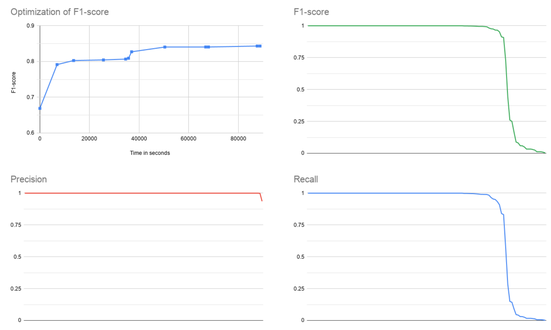

Figure 3. Optimization of the F1-score and results of precision and recall analysis

Software implementation

The cleaning of the data and pattern extraction are implemented in Microsoft SQL Server Management. In Python, the construction of rules and string similarities, clustering and optimization are performed based on equations (1-3).

Conclusions

A disambiguation method for entity resolution in databases is proposed. To optimize the method, precision and recall analysis is performed using a golden set of clusters and the F1-score is maximized using simulated annealing. The research method is used to disambiguate scientific references and create clusters of records that refer to the same bibliographic entity. The method starts by pre-cleaning the records and extracting bibliographic labels. Next, we construct rules based on these labels and make use of the tf-idf algorithm to compute a string similarity measure. We create clusters by means of a rule-based scoring system. Finally, we perform precision and recall analysis using a golden set of clusters and optimize our parameters with a simulated annealing algorithm. Here we show that it is possible to optimize the precision and recall of a disambiguation method using a global optimization algorithm. The proposed method is generic and applicable on similar entity resolution problems.

References

[1] Zhao K., Caron E. and Guner S., Large scale disambiguation of scientific references in patent databases, Proceedings of STI, 2016.

[2] Nijland R., The Disambiguation of Scientific References in the PatStat Database, MA thesis, Tilburg University, 2017.

[3] Patstat, Epo.org, Worldwide Patent Statistical Database, 2020.