An Optimization Method for Entity Resolution in Databases: With a Case Study on the Cleaning of Scientific References in Patent Databases

Emiel Caron
Department of Management, Tilburg University, Warandelaan 2, Tilburg, The Netherlands

Keywords: Entity Resolution, Data Disambiguation, Data Cleaning, Data Integration, Bibliographic Databases.

Abstract: Many databases contain ambiguous and unstructured data which makes the information it contains difficult to use for further analysis. 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 propose a generic optimization method for disambiguating large databases. This method is used on a table with scientific references from the Patstat database. The table holds ambiguous information on citations to scientific references. The research method described is used to 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 string similarities. We create clusters by means of a rule-based scoring system. Finally, we perform precision-recall analysis using a golden set of clusters and optimize our parameters with simulated annealing. Here we show that it is possible to optimize the performance of a disambiguation method using a global optimization algorithm.

1 INTRODUCTION

Ambiguous data in a data set can lead to misleading, inaccurate and incomplete results in analysis. The problem of ambiguous and unstructured data in large-scale databases is a common problem (Bhattacharya and Getoor, 2007). This is also known as the *record linkage problem* and *entity resolution* - the general term for disambiguating records that refer to the same entity by linking and classifying them as one group. Since manual disambiguation is not attainable in large databases, there is need for automatic methods. In the past, research has been conducted on finding automatic methods for the disambiguation of records in databases. For instance, Caron and Van Eck (2014) presented an author name disambiguation method for large databases, and Caron and Daniels (2016) provided a method for identifying company name variants in large databases.

In this work we want to deal with two problems in the current entity resolution approaches. The first problem is how to develop a generic method for disambiguating large databases that, in theory, is applicable to multiple domains. The second problem is how to optimize the parameters of a generic method. The optimization of the method refers to finding the model's parameters that provide the best results, i.e. the highest value of a certain performance measure, typically precision-recall metrics. To the best of our knowledge, none of the existing approaches for entity resolution in databases use heuristics to optimize performance. By making the method more efficient and by implementing optimization techniques, we aim to reduce computation time and obtain better results. This in turn would open up the possibility for other data leaning methods to be optimized as well.

1.1 Entity Resolution

In this paper, data disambiguation, record linkage and entity resolution all refer to the same topic: we want to determine a mapping from ambiguous database records to real-world entities. In the last 20 years much research has been conducted on this topic. Here we analyze a subset of past research that is relevant for this paper. Furthermore, we briefly identify particular methods and techniques used to disambiguate patent databases. Entity resolution is the problem of identifying records that refer to the same real-world entity. For example, the names 'J. Doe' and 'Doe, Jon' may actually refer to the same person. Without unique identifiers, this so-called

Caron, E.
An Optimization Method for Entity Resolution in Databases: With a Case Study on the Cleaning of Scientific References in Patent Databases.
DOI: 10.5220/0009972406250632
In *Proceedings of the 15th International Conference on Software Technologies (ICSOFT 2020)*, pages 625-632
ISBN: 978-989-758-443-5

duplication problem can lead to various issues such as erroneous computed statistics, redundant data and data inconsistencies (Monge and Elkan, 1996; Hernández and Stolfo, 1995). In the context of databases, Bhattacharya and Getoor (2007) state that the entity resolution problem includes (a) identifying the underlying entities of a database and (b) labelling records of the database with the entities they refer to. Typically, these two issues cannot be solved independently. Bhattacharya and Getoor (2007) describe two main problems on why entity resolution is difficult to solve. The first is the identification problem which occurs when entities are referred to by different name variants in a database. Second, there is a disambiguation problem when two similar records actually do not refer to the same entity. Failure in disambiguation leads to lower precision; the identification problem affects the recall of a cleaning method.

This paper is organised as follows. Related work is presented in Section 2. In Section 3, our method for entity resolution is explained step-by-step. After that the method is applied on a case study involving disambiguation of scientific references in the Patstat database in Section 4. Here the method's performance is evaluated with precision-recall analysis. Finally, we discuss conclusions and further research in Section 5.

2 RELATED WORK

There are different existing approaches to entity resolution. The traditional approach is attribute-based, where an attribute refers to a property or characteristic of a record (Fellegi and Sunter, 1969). In these approaches a similarity value is computed for every pair of records based on their attributes. If a pair's similarity is above a certain threshold, the records are considered to refer to the same entity. Different similarity measures may be used for computing a similarity value.

In general, manual disambiguation is not feasible in large databases. Therefore, computerized methods are required. From the class of automated data cleaning methods, we first review general data mining methods, and after that we review specific work on the large-scale cleaning of bibliographic databases like Patstat, directly related to the case study in Section 4.

In supervised approaches, a data set with labelled records is used to train the learning scheme. However, large data sets with manually labelled records would be expensive to collect and are often not available. For this reason research has focused on developing unsupervised approaches for entity resolution. Unsupervised approaches use similarity metrics and clustering algorithms to find clusters of name variants. While unsupervised approaches do not require training data sets, they often perform less well than supervised approaches (Levin et al., 2012).

Benjelloun et al. (2009) present a generic entity resolution method and algorithms based on pairwise decisions. Pairwise means that the method matches two records at a time. In addition, they elaborate on the formal properties of an efficient algorithm. However, no solution is given to find the optimal settings for the algorithm. This paper addresses this issue.

Nguyen and Ichise (2016) describe an effective supervised solution to entity resolution named cLink for linking records from different data sources. cLink is a generic method that uses a learning algorithm to optimize the configuration of matching properties and similarity measures. To this end, the learning algorithm cLearn requires a training data set to compute the F1-score of the candidates with the highest matching score. The optimized configuration of similarity functions is used to compute a final matching score for every candidate pair. Only pairs with a maximal matching score that is above a certain threshold are considered to be co-referent. Related to the work of Nguyen and Ichise (2016) are the active learning approaches RAVEN (Ngomo et al., 2011), EAGLE (Ngonga Ngomo and Lyko, 2012) and ActiveGenLink (Isele and Bizer, 2013). These three systems all work similarly by using active learning to achieve accurate link specifications. Active learning requires interaction with the user to confirm or decline a certain match. EAGLE and ActiveGenLink additionally apply genetic programming to achieve better efficiency.

Thoma and Torrisi (2007) compare two approaches to automatic matching techniques on linking patent data from the PatStat database to data on firms from the Amadeus database. Amadeus is a financial database on European firms, maintained by Bureau Van Dijk. Both approaches start with name standardization which consists of tasks such as punctuation cleaning, character cleaning and removal of common terms such as 'company' and 'co.'. Next, the first approach uses character-to-character comparison (perfect matching). The second approach uses approximate string comparison based on string similarity functions (approximate matching). To measure string similarity Thoma and Torrisi make use of the Jaccard index combined with tf-idf weights. Even though perfect matching yields high precision, the number of matches can be quite low. Approximate matching can increase the number of matches, but at the cost of pre-

cision. The results of the two approaches show that approximate matching indeed yields a significant increase in the number of matches, while producing a limited loss of precision. Thoma et al. (2010) developed a matching technique that, in addition to name similarity, compares other information such as location and founding year to match firms in multiple large databases. This technique uses a combination of dictionary-based methods and rule-based approaches. Dictionary-based methods depend on large data sets with verified name variants. While these methods are simple and very precise, typically the number of matches is low in the case of firm names. Rule-based approaches set up a collection of rules to determine the similarity between different firm names. Thoma et al. focus on approximate matching in their set of rules. Following Thoma et al., Lotti and Marin (2013) describe an improved cleaning routine to match applicants in the PatStat database to firms in the AIDA database. AIDA is a product of Bureau Van Dijk containing information on Italian companies. In addition to name harmonization and matching, Lotti and Marin (2013) use harmonized addresses for exact matching of firms. Furthermore, approximate matches on names were used in combination with visual checks. Finally, the name variants contained in the data set created by Thoma et al. (2010) were used for further matching.

3 METHOD FOR DATA DISAMBIGUATION

When studying past research on entity resolution, specifically the work described in (Caron and Van Eck, 2014; Zhao et al., 2016; Caron and Daniels, 2016), we observe that the methods on cleaning databases all have similar components and approaches. This common approach is in essence as follows. Features are extracted from the records in the database and similarity rules are constructed on the basis of these features. These rules provide evidence that a pair of records match. Next, scores are assigned to every rule and the total score for all pairs is computed. These scores measure the similarity of a pair of records. The pairs that score above a certain threshold are linked using a clustering algorithm, e.g. connected components or maximal cliques. After that the performance of the clusters is benchmarked against a golden verified set with precision-recall analysis. Subsequently, the scores and threshold parameters are updated experimentally to improve the average values for both precision and recall. We add to this known approach a way to optimize the method's parameters with a global optimization algorithm to obtain the clusters with the highest possible average precision and recall.

3.1 Generic Method

Here the generic method is presented. Consider N objects or representations of entities $r_1,...,r_N$. For $i = 1,...,N$, the object r_i has feature vector f^i consisting of M features $f^i = (f^i(1),...,f^i(M))$. Every feature $f^i(j)$ has a domain F_j, i.e. $f^i(1) \in F_1,...,f^i(M) \in F_M$ which is independent of i.

We call two entities *similar* if their corresponding feature vectors are similar. Stated more precisely, we define

$$\sigma_l : F_l \times F_l \to \mathbb{R}^+ \cup \{0\} \tag{1}$$

such that $\sigma_l(f^i(l), f^j(l))$ measures the similarity of any pair $(f^i(l), f^j(l)) \in F_l \times F_l$, where $i \neq j$.

To obtain a similarity score as a single number, we define a NxN matrix S with elements s_{ij} and weight vector $w = (w_1,...,w_M)$:

$$s_{ij} = \sum_{k=1}^{M} w_k \cdot \sigma_k(f^i(k), f^j(k)) \quad \text{where } w_k \geq 0, i \neq j \tag{2}$$

Objects that are similar are grouped into sets as follows. Suppose we have Q sets $\Sigma_1,...,\Sigma_Q$. Define a threshold δ. Then

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

where

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

unless $|\Sigma_p| = 1$.

The sets $\Sigma_1,...,\Sigma_Q$ are mutually exclusive:

$$\Sigma_p \cap \Sigma_q = \emptyset \quad \text{for } p \neq q$$

and are chosen to be maximal:

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

The method's performance is evaluated using precision and recall analysis. The F1-score is the harmonic mean of precision and recall (Fawcett, 2006). Since the objective is to obtain clusters with both high precision and high recall, we maximize the F1-score. Therefore we choose our parameters $w = (w_1,...,w_M)$ and δ in such a way that

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

where our objective function $L(w, \delta)$ is the average F1-score of the clusters. Optimizing the method's parameters is done using a simulated annealing algorithm (Xiang et al., 1997).

3.2 Optimization

The objective function in the optimization problem described in Eq. (3) is nonlinear and yields many local optima. Moreover, according to Erber and Hockney (1995), the number of local optima typically increases exponentially as the number of variables increases. Here the variables are weights w and thresholds δ used in the method described in section 3.1. We need to use a global optimization method such as the simulated annealing algorithm or a genetic algorithm in order to find a global optimum instead of getting trapped in one of the many local optima.

The problem of maximizing the F1-score with respect to the weights w and thresholds δ of the model is a combinatorial optimization problem. Research in combinatorial optimization focuses on developing efficient techniques to minimize or maximize a function of many independent variables. Since solving such optimization problems exactly would require a large amount of computational power, heuristic methods are used to find an approximate optimal solution. Heuristic methods are typically based on a 'divide-and-conquer' strategy or an iterative improvement strategy. The algorithms we focus on are based on iterative improvement. That is, the system starts in a known configuration of the variables. Then some rearrangement operation is applied until a configuration is found that yields a better value of the objective function. This configuration then becomes the new configuration of the system and this process is repeated until no further improvements are found. Since this method only accepts new configurations that improve the objective function, the system is likely to be trapped in a local optima. This is where simulated annealing plays its part.

Simulated annealing is inspired by techniques of statistical mechanics which describe the behavior of physical systems with many degrees of freedom. The simulated annealing process starts by optimizing the system at a high temperature such that rearrangements of parameters causing large changes in the objective function are made. The "temperature", or in general the control parameter, is then lowered in slow stages until the system freezes and no more changes occur. This cooling process ensures that smaller changes in the objective functions are made at lower temperatures. The probability of accepting a configuration that leads to a worse solution is lowered as the temperature decreases (Kirkpatrick et al., 1983). To optimize the F1-score of our method we use the dual_annealing() function of the SciPy.optimize package in Python (SciPy.org, 2020). This dual annealing optimization is derived from the research of Xiang et al. (1997) in combination with a local search strategy (Xiang and Gong, 2000).

Next to simulated annealing, the genetic algorithm was explored for finding a global optimum. The genetic algorithm computes a population and produces generations by combining and altering individual solutions of the population (Whitley, 1994). While the genetic algorithm is often used in optimization problems, for our problem simulated annealing seems to be a better fit. The genetic algorithm requires large generations of populations to be computed which would be very time consuming considering our problem as the computation of one individual already takes several minutes.

3.3 Method Steps

In Figure 4, an overview is given of the practical steps in the method for entity resolution detailed in the previous section. The input of the method is raw data that corresponds to ambiguous name variants and the output are clusters of name variants. The method is composed out of main 5 steps:

1. *Pre-processing*.
2. *Filtering*.
3. *Rule-based clustering*.
4. *Post-processing*.
5. *Optimization*.

Step 1 involves pre-cleaning the data and extracting and evaluating descriptive labels. In step 2, the tf-idf algorithm (Salton and Buckley, 1988) is used to compute string similarities between records using Eq. (1). By constructing rules based on the extracted labels, evidence is collected for the similarity between records and based on that candidate record pairs are created. In step 3, rules are scored and combined using Eq. (2) and after that the connected-components algorithm (Bondy and Murty, 1976) is used to obtain clusters of records. In step 4, records for which no duplicates are identified are assigned to new single-record clusters. In step 5, the clusters are evaluated on the golden sample using precision and recall analysis. The parameters used for constructing and scoring rules, such as thresholds and scores, are adjusted to achieve higher values in precision and recall. Using simulated annealing (Xiang et al., 1997) we obtain the optimal parameters. Here the average F1-score over all clusters is used, the harmonic mean of the precision and recall measure, as our objective function (Fawcett, 2006) as defined in Eq. (3). The method stops when the global optimum is obtained.

3.4 Software Implementation

The generic method as described in section 3.1 is implemented in Python[1]. In the example case of the Patstat table with ambiguous scientific references, the objects $r_1, ...r_N$ are the records of the table and the features of each record are described in the table 'evaluated_patterns'. The features in the case study Section 4 are extracted bibliographic meta information as: publication title and year, author names, journal information, etc. Eq. (1) of the generic method refers to the rules that are constructed. These rules provide evidence that two records are similar. In order to compute the string similarities for the rules we use an efficient implementation of tf-idf in Python. The scores that are assigned to the rules correspond to weight vector w in Eq. (2). The clustering of records is done using the connected components algorithm and results in the sets $\Sigma_1, ...\Sigma_Q$ as described in the generic method.

The method '`find_clusters()`', given in the script '`find_clusters.py`', implements the whole generic method and computes the F1-score. Its input parameters are a configuration of variables (w, δ) and a table containing feature vectors. We use this method as our objective function for the `dual_annealing()` method (SciPy.org, 2020) described in the script '`optimize.py`'. In this way we find the optimal configuration of the parameters in Eq. (3).

4 CASE STUDY ON PATSTAT DATA

4.1 Background

PatStat (European Patent Office, 2019) is a worldwide patent database from the European Patent Office. It contains bibliographic information on patent applications and publications and is important source in the field of patent statistics. One of its tables, TLS214, holds information on scientific references cited by patents. These references are collected from patent application, in which patent applicants reference (scientific) literature to acknowledge the contribution of other writers and researchers to their work. In 2019, this table contained more than 40 million records with scientific references. Therefore, the table is an important point of reference for research, for example, that studies the relation between science and technology. However, amongst the scientific references there are many name variants of publications caused by missing data, inconsistent input convention, different order of items, typos, etc. The records in Figure 1 serve as an example of such name variants. In order for table TLS214 to be a reliable point of reference for research, its records need to be disambiguated and restructured (Zhao et al., 2016).

[1]https://emielcaron.nl/wp-content/uploads/2020/05/entity_resolution_optimization_code.7z

4.2 Results

Here we present and analyze the results of applying the disambiguation method on scientific references from the Patstat database, by comparing the set of clusters generated by the method with a golden set of 100 clusters. The golden sample is a sample of records with verified clusters (Caron and Van Eck, 2014), evaluated by human domain experts. We run our algorithm and keep track of the intermediate results. After approximately 24 hours, no more improvements are made to the F1-score, so the algorithm is stopped. The plot in Figure 2 shows the increase of the F1-score against time in seconds. The first point of the line is the F1-score of the initial clusters. It is important to note that, this plot shows that the simulated annealing algorithm indeed is able to improve the F1-score and thus our method produces better clusters. We observe that the algorithm makes large improvements to the objective value in the beginning and smaller improvements towards the end. This is in accordance with the theory on simulated annealing, as the smaller changes are made at low temperatures near the end.

The results of our method, depicted in Table 1, are promising in terms of precision-recall statistics. The average values of precision, recall, and F1 are high for all cluster categories. In comparison, the clusters achieve significantly better precision and recall values than the initial clusters created with the method's initial parameters, as depicted in development of the F1-score in Figure 2. The plots in Figure 5 show a large positive shift to the right in the distribution of the different measures, compared to initial values. We also

npl_publn_id	npl_biblio
971443839	LOWRY OH; ROSEBROUGH NJ; FARR AL; RANDALL RJ. 'Protein measurement with the Folin phenol reagent' J BIOL CHEM. vol. 193, no. 1, 1951, pages 265 - 75
971346160	Lowry O H. et al., Protein measurement with the folin phenol reagent, The Journal of Biological Chemistry, 1951, vol. 193, pp. 265-275.
971334489	LOWRY OH; ROSEBROUGH NJ; FARR AL; RANDALL RJ. ' Protein measurement with the Folin phenol reagent' J BIOL CHEM. vol. 193, no. 1, 1951, pages 265 - 75
971202021	Lowry, O.H. et al., Protein measurement with the folin phenol reagent, J. Biol. Chem., 1951, vol. 193, pp. 265-275.
971250368	LOWRY OH ET AL.: 'Protein measurement with the Folin phenol reagent' JBIOL CHEM vol. 193, 1951, pages 265 - 275
971020960	LOWRY OH; ROSEBROUGH NJ; FARR AL; RANDALL RJ. 'Protein measurement with the Folin phenol reagent' J BIOL CHEM. vol. 193, no. 1, 1951, pages 265 - 75
969668224	Lowry, O H., et al., Protein measurement with the Folin phenol reagent, J Biol Chem., 193(1): 265-275, (Nov. 1, 1951).
969666389	Lowry et al., Protein Measurement with the Folin Phenol Reagent, Biological Chemistry, 193, pp. 265-275, 1951, www.jbc.org.
969909223	Lowry et al., Protein Measurement with the folin phenol reagent" Dept. of Pharma., Washington Univ. School of Med. 265-275 (1951).
969981521	O.H. LOWRY, N.J. ROSEBROUGH, A.E. FARR, R.J. RANDALL: 'Protein measurement with Folin phenol reagent' J. BIOL. CHEM. vol. 193, 1951, pages 265 - 275
969522409	LOWRY OH ET AL.: 'Protein measurement with the Folin phenol reagent' JBIOL CHEM vol. 193, 1951, pages 265 - 275
969558715	LOWRY, O.H., N.J. ROSEBROUGH, A.L. FARR, R.J. RANDALL: 'Protein measurement with the Folin phenol reagent' J. BIOL CHEM vol. 193, 1951, page 265
969459223	Lowry, O.H., et al., Protein Measurement with the Folin Phenol Reagent., J. Biol. Chem., 1951, vol. 193, pp. 265-75.
969441150	LOWRY, O. H. ET AL.: 'Protein measurement with the Folin phenol reagent' J BIOL CHEM vol. 193, no. 1, 1951, pages 265 - 75
969398320	Lowry et al., Protein Measurement with the Folin Phenol Reagent, J. Bio. Chem., 1951, vol. 193, pp. 265-275.
969134625	Oliver H. Lowry, et al., Protein Measurement with the Folin Phenol Reagent, The Journal of Biological Chemistry,193, May 28, 1951, pp. 265-275.
969148285	LOWRY, O.H., N.J. ROSEBROUGH, A.L. FARR, R.J. RANDALL: 'Protein measurement with the Folin phenol reagent' J. BIOL CHEM vol. 193, 19 March 0004, page 265
969148288	LOWRY, O.H., N.J. ROSEBROUGH, A.L. FARR, R.J. RANDALL: 'Protein measurement with the Folin phenol reagent' J. BIOL CHEM vol. 193, 19 March 0004, page 265
970708951	LOWRY, O.H.; ROSEBROUGH, N.J.; FARR, A.L.; RANDALL, R.J.: 'Protein measurement with the Folin phenol reagent.' J BIOL CHEM vol. 193, 1951, pages 265 - 275
970598256	Lowry, O.H., et al., Protein Measurement With the Folin Phenol Reagent., J. Biol. Chem., 1951, 193, 265.
970663260	Lowry, et al., Protein Measurement with the Folin Phenol Reagent, J. Biol. Chem. 193: 265-275(1951).

Figure 1: A set of name variants that refer to the article 'Protein measurement with the Folin phenol reagent'. A selection of records from a cluster in the golden sample.

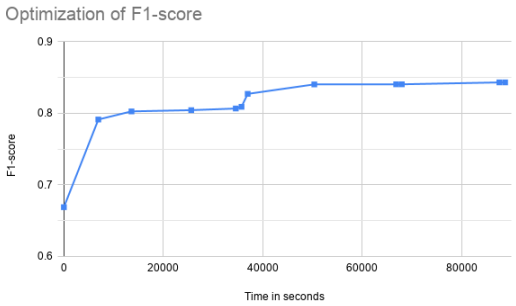

Figure 2: Optimization of the F1-score against time in seconds.

see that the F1-score almost solely depends on the recall measure as precision is equal to 1 for almost all clusters.

Table 1: Statistics of optimized clusters.

Category	# of clusters	Precision (Avg)	Recall (Avg)	F1 (Avg)
Large	31	0.9979	0.9893	0.9933
Medium	43	1	0.9490	0.9563
Small	41	1	0.5936	0.6114
Total	115	0.9994	0.8331	0.8433

Figure 6 shows an example cluster of correctly classified scientific references. We observe that the optimized method correctly identifies all records of the cluster and achieves a precision and recall value of 1.

In addition, when analyzing the clusters, we notice that our method is slightly conservative, it values precision over recall. As a result the method is more likely to create multiple clusters with high precision than to create one large cluster with potential errors. Typically, the method splits the name variants of one golden cluster into one large dominant cluster and multiple small clusters. This is illustrated for golden cluster 100 in Figure 3. The number inside

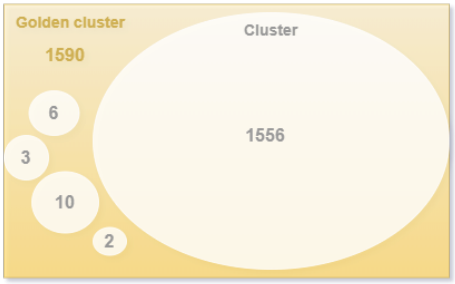

Figure 3: Illustration of a golden cluster distributed over multiple system clusters.

each cluster is its size, i.e. the number of records in the cluster. The remaining records of the golden cluster were not paired by our method or did not score above the threshold and were therefore not clustered.

The sizes of the golden cluster range between 2 and 1590 records. Because of this variety, we distinguish between three cluster categories. Large clusters contain more than 100 records and typically score high on recall. These clusters contain the majority, or all, of the records of the corresponding golden cluster. Medium clusters contain 11 to 100 records and typically correspond to medium-sized golden clusters. Most small clusters, containing 10 or less records, are clusters like the small clusters in figure 8. Because of this, recall is sometimes lower for small clusters.

5 CONCLUSIONS

Here an optimization 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 average F1-score of disambiguation method using a global optimization algorithm. The proposed method is generic and applicable on similar entity resolution problems.

The results of this research are beneficial for both academics and industry. Academics can use our method to efficiently access information in an ambiguous data sets. The unambiguated data in Patstat can for example be used to study the effect of science outputs on industry patents. Data scientists at firms often have to deal with ambiguous data. The disambiguation method studied provides a way to efficiently clean such data sets so that these can be used for data analysis. To the best of our knowledge, none of the existing approaches for entity resolution in bibliographic databases use heuristics to optimize the parameters for rule construction and clustering. Hence, the results of this research provide new insights for research on entity resolution.

In future work, we want to compare our method with existing methods on benchmark datasets for entity resolution. In addition, we could speed up the method by incorporating parallel computing. Especially the construction of rules is suitable for parallel

computing as the rules could be split over multiple cores of a computer and be constructed simultaneously. This would decrease the computation time of the disambiguation method significantly. Moreover, the exploration of different optimization techniques for finding a global optimum next to simulated annealing would be very interesting for comparison. A potential candidate for comparison is Tabu search.

ACKNOWLEDGEMENTS

We kindly acknowledge Wen Xin Lin and Prof. H.A.M. Daniels for their contribution to this work.

REFERENCES

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., and Widom, J. (2009). Swoosh: A generic approach to entity resolution. *The VLDB Journal*, 18(1):255276.

Bhattacharya, I. and Getoor, L. (2007). Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 1(1).

Bondy, J. A. and Murty, U. S. R. (1976). *Graph theory with applications*, volume 290. Macmillan London.

Caron, E. and Daniels, H. (2016). Identification of organization name variants in large databases using rule-based scoring and clustering - with a case study on the web of science database. In *Proceedings of the 18th International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 182–187. INSTICC, SciTePress.

Caron, E. and Van Eck, N.-J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In Noyons, E., editor, *Proceedings of the Science and Technology Indicators Conference 2014*, pages 79–86. Universiteit Leiden.

Erber, T. and Hockney, G. (1995). Comment on method of constrained global optimization. *Physical review letters*, 74(8):1482.

European Patent Office (2019). *Data Catalog - PATSTAT Global*, 2019 autumn edition edition.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Hernández, M. A. and Stolfo, S. J. (1995). The merge/purge problem for large databases. In *ACM Sigmod Record*, volume 24, pages 127–138. ACM.

Isele, R. and Bizer, C. (2013). Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web*, 23:2–15.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Levin, M., Krawczyk, S., Bethard, S., and Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, 63(5):1030–1047.

Lotti, F. and Marin, G. (2013). Matching of patstat applications to aida firms: discussion of the methodology and results. *Bank of Italy Occasional Paper*, (166).

Monge, A. E. and Elkan, C. (1996). The field matching problem: Algorithms and applications. In *KDD*, volume 2, pages 267–270.

Ngomo, A.-C. N., Lehmann, J., Auer, S., and Höffner, K. (2011). Raven–active learning of link specifications. *Ontology Matching*, 2011.

Ngonga Ngomo, A.-C. and Lyko, K. (2012). Eagle: Efficient active learning of link specifications using genetic programming. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications*, pages 149–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nguyen, K. and Ichise, R. (2016). Linked data entity resolution system enhanced by configuration learning algorithm. *IEICE Transactions on Information and Systems*, E99.D(6):1521–1530.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523.

SciPy.org (2020). Scipy.optimize package with dual_annealing() function.

Thoma, G. and Torrisi, S. (2007). *Creating powerful indicators for innovation studies with approximate matching algorithms: a test based on PATSTAT and Amadeus databases*. Università commerciale Luigi Bocconi.

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., and Harhoff, D. (2010). Harmonizing and combining large datasets-an application to firm-level patent and accounting data. Technical report, National Bureau of Economic Research.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85.

Xiang, Y. and Gong, X. (2000). Efficiency of generalized simulated annealing. *Physical Review E*, 62(3):4473.

Xiang, Y., Sun, D., Fan, W., and Gong, X. (1997). Generalized simulated annealing algorithm and its application to the thomson model. *Physics Letters A*, 233(3):216–220.

Zhao, K., Caron, E., and Guner, S. (2016). Large scale disambiguation of scientific references in patent databases. In Rafols, I., Molas-Gallart, J., Castro-Martinez, E., and Woolley, R., editors, *Proceedings of 21st International Conference on Science and Technology Indicators (STI 2016)*, pages 1404–1410. Editorial Universitat Politècnica de València.

APPENDIX

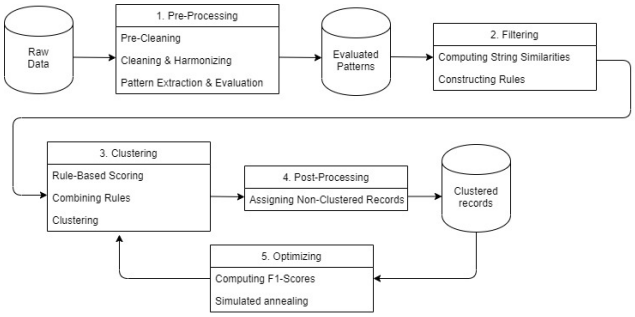

Figure 4: Overview of the disambiguation method.

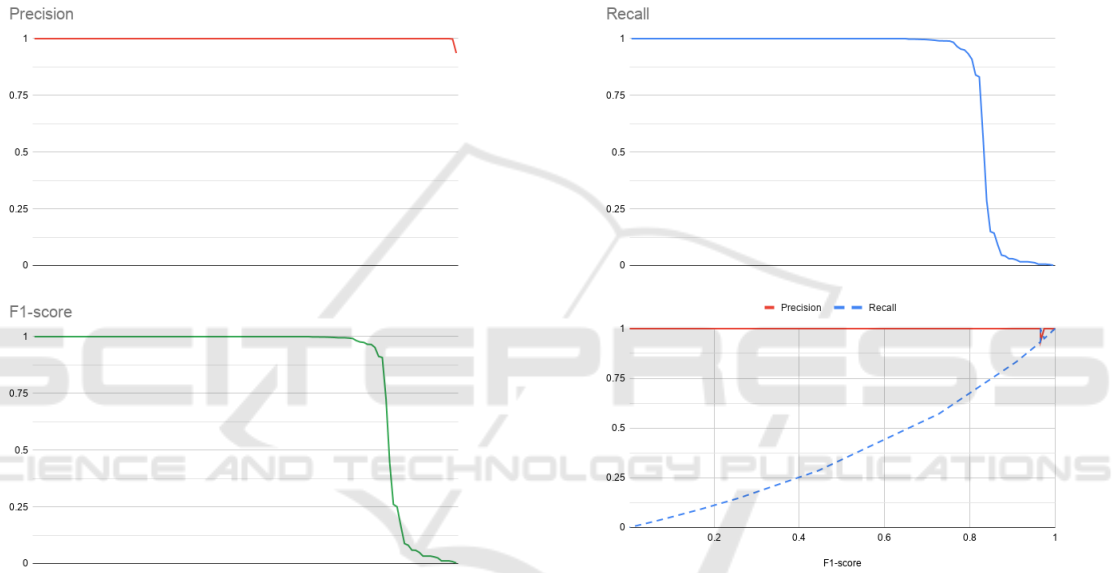

Figure 5: Distribution of the precision and recall measure and the F1-score on all optimized clusters. The clusters on the x-axis are ranked based on precision-recall-f1 values.

cluster_id	npl_publn_id	npl_biblio
		B., Jun. 15, 1976, vol. 13, No. 12, pp. 5188-5192.
96	959522198	G. Kresse et al., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Physical Review B, vol. 54, No. 16, Oct. 15, 1996, pp. 169-186.
96	969333849	Kresse et al., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set Physical Review B, vol. 54, No. 16, pp. 11169-11186, (Oct. 15, 1996).
96	966302253	Kresse, Efficient Iterative Schemes for ab initio Total-Energy Calculations Using a Plane-Wave Basis Set, Physical Review B, vol. 54, No. 16, pp. 11169-11186, (Oct. 1996).
96	968776323	Kresse et al., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B., vol. 54, No. 16, pp. 11169-11186 (1996).
96	971457332	Kresse, G., Furthmüller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set Phys. Rev. B 54, 11169-11186 (1996)
96	967055152	Kresse, G., Fürthmuller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Physical Review B, vol. 54, No. 16, pp. 11169-11186.
96	959773698	G. Kresse, J. Furthmuller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B54(16), pp. 11169-11186 (1996).
96	959773797	G. Kresse, J. Furthmuller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B54(16), pp. 11169-11186 (1996).
96	959773896	G. Kresse, J. Furthmuller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B54(16), pp. 11169-11186 (1996).

Figure 6: Contents of example cluster 96.