Identification of Organization Name Variants in Large Databases using Rule-based Scoring and Clustering

With a Case Study on the Web of Science Database

Emiel Caron[1] and Hennie Daniels[1,2]

[1]*Erasmus Research Institute of Management, Erasmus University Rotterdam, P.O. Box 1738, Rotterdam, The Netherlands*

[2]*Center for Economic Research, Tilburg University, P.O. Box 90153, Tilburg, The Netherlands*

Keywords: Large Scale Databases, Data Warehousing, Database Integration, Data Cleaning, Data Mining, Clustering.

Abstract: This research describes a general method to automatically clean organizational and business names variants within large databases, such as: patent databases, bibliographic databases, databases in business information systems, or any other database containing organisational name variants. The method clusters name variants of organizations based on similarities of their associated meta-data, like, for example, postal code and email domain data. The method is divided into a rule-based scoring system and a clustering system. The method is tested on the cleaning of research organisations in the Web of Science database for the purpose of bibliometric analysis and scientific performance evaluation. The results of the clustering are evaluated with metrics such as precision and recall analysis on a verified data set. The evaluation shows that our method performs well and is conservative, it values precision over recall, with on average 95% precision and 80% recall for clusters.

1 INTRODUCTION

In many databases, one organisational entity is listed in the records with many associated name variants. For example, the Leiden University (2015) has many name variants in the Web of Science (WoS) database, like: University Leiden, Leiden Universiteit, Leiden State University, State University Leiden, Leiden University Hospital, State University Leiden Hospital, Leiden Universitair Medisch Centrum, LUMC, and so on. Large companies often have many name variants, e.g. the technology company Royal Philips has several hundreds of name variants in the Patstat (2015) database, which is a statistical database filled with patent information. Obviously, manual normalisation of organisation names is not feasible in large databases, which might list millions of companies and organisations.

The research problem that is addressed here is: "How can organization name variants be identified automatically in large databases?". The answer to this problem is given by a general method for the identification of organization name variants using rule-based scoring and clustering proposed in this paper. The method is able to cluster name variants in large databases with millions of records in an efficient way. The emphasis of the method is on the cleaning of names not on unification. The results of this method are useful for any analysis involving correct and unified organisation names, such as: company patent analysis, evaluative bibliometrics and the ranking of scientific institutes, the assessment of cooperation and communication between organizations, and the creation of linkages, based on company names, between Customer Relationship Management databases.

Data cleaning is often the necessary step prior to knowledge discovery and business analytics. Automatic data cleaning methods can be categorised in several groups (Maletic and Marcus, 2010): transformational rules, statistical methods for numeric data, and data mining methods, such as cluster and pattern matching techniques (Cohen et al., 2003, Koudas et al., 2004, Morillo et al., 2013), for categorical data. Data mining methods for the identification of organisation name variants and person name disambiguation are divided into supervised and unsupervised learning approaches. In supervised learning approaches, a classifier is trained on a data set with pairs of records, where organisation with similar names are classified as being the same entity or a different entity. The problem with

Caron, E. and Daniels, H.
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 (ICEIS 2016) - Volume 1*, pages 182-187
ISBN: 978-989-758-187-8

supervised approaches in this context, is that a large, manually checked, representative, data set is required for training. Such a data set is usually not available, which makes supervised approaches for our problem hard to use. In unsupervised learning, a metric of similarity is defined between pairs of records, that describe an organisational entity, and after that a clustering algorithm is applied (Levin et al., 2012, Song et al., 2007). The method described in this paper is based on unsupervised rule-based clustering, in combination with approximate string pattern matching. A clear advantage of our method is that the matching rules are easy to understand and combine.

The organization of this paper is as follows. In the next section, the phases of the method for the clustering of organisation names are explained in detail. After that the method is evaluated, with precision-recall analysis, on the WoS database for the clustering of scientific organisation names. We close the paper with some concluding remarks and proposals for further research.

2 METHODOLOGY

A visual summary of the process for the identification of organisation name variants is provided by Figure 1 in the Appendix. The method is composed out of three stages:

1. Pre-processing;
2. Rule-based scoring and clustering;
3. Post-processing.

Organisational meta-data from a source database is taken as input in the process and clusters of organisation name variants are produced as output. Typical examples of important meta-data available for organisation name matching are: country, city, postal code, street, organisation type, email domain, etc. The method is designed to cluster all organisation name variants in the whole database. In the case study the method is applied on the WoS database (version April 2013) with roughly 124 million publication records. Moreover, the method is implemented with a combination of Microsoft SQL Server and Visual Studio, where SQL server is used for the data handling and Visual Studio for the implementation of the cluster algorithm.

2.1 Pre-processing

In the pre-processing stage the relevant meta-data items are cleaned and harmonized to improve the data quality, and helper tables are created for the subsequent clustering stage.

Postal code data is cleaned an put into a consistent format. Besides, postal codes are classified into groups, indicating the number of different organisations present in a postal code area. Groups with a relative high number of organisations are treated under a stricter regime, e.g. with a higher threshold.

From the available data, the organisational types are determined, such as ‘company’, ‘bank’, ‘university’, ‘hospital’, ‘institute’, etc., with string extraction patterns and regular expressions.

An important data element for clustering, when it is available, is the email domain address that is linked with an organisation, because it is very discriminative. Usually, multiple email domains are connect to large organisations, e.g. Leiden University uses ‘leidenuniv.nl’, but also ‘liacs.nl’ and ‘lumc.nl’. In the pre-processing stage, the email domains are replaced by their most popular or recent variant. Email domains that cannot be directly linked to an organisation, e.g. ‘gmail.com’ or ‘hotmail.com’, are removed, because they cannot be used in an meaningful way.

2.2 Rule-based Scoring and Clustering

In this stage, the clusters are created that identify likely name variations of the same organization, in the following steps (see Figure 1):

a. A set of rules is created that produces pairs of organizational names with common characteristics and string name similarity. These rules target specific elements of the organizations’ characteristics such as combinations of the country, city, postal code, email address, organisation type.
b. A scoring system is applied on the rules and scores are computed for created record pairs.
c. Pairs above the threshold are clustered with an algorithm, taking linear time, that searches for connected components.

2.2.1 Apply Rules (Step 2a)

In step 2a, the objective is to create pairs of organisation name variants by self-joining the tables that result from the pre-processing stage. For this purpose scoring rules are used that are depicted in Table 1, where rules 1-4 are the basic rules and rules 5-9 are combinations of the basic elements. The score for a rule increases when it contains more meta-data elements, indicating that there is more proof that a record pair is a name variant of each other. Therefore

rule 9 is considered a stronger rule than rule 1. The number of rules and scores can easily be adapted if more relevant meta-data is available. The rule score values and the threshold values are based on domain expert knowledge about the database under consideration. The values are fine-tuned in the initial evaluation of the method on a verified data set.

The scoring system assigns scores to record pairs from the strongest to the weakest rules. If a record pair is scored on a strong rule, the pair is not considered for the weaker rules, to prevent additional scoring. Not listed in Table 1, is that there are additional constraints for each rule, like the size of the city, the type of postal code (general or specific), and so on, that are configured very strict for strong rules but are given more degrees of freedom for weak rules.

Table 1: Example rules with associated meta-data, organisation name similarity, and rule scores. The threshold value is 4 in the example.

Rules	Country	City	Postal code	Email	Org. type	Name sim.	Score
Rule 1	✓	✓	✓				1
Rule 2	✓	✓		✓			1
Rule 3	✓	✓			✓		1
Rule 4	✓	✓				✓	2
Rule 5	✓	✓	✓	✓			2
Rule 6	✓	✓		✓	✓		2
Rule 7	✓	✓	✓		✓		2
Rule 8	✓	✓	✓	✓	✓		4
Rule 9	✓	✓	✓	✓	✓	✓	10

Notice that rules always only hold in a specific country and city, because organisation names might not be unique in the whole world or even in a specific country. The rules use the meta-date: postal code information (rule 1), email domain (rule 2), organisation type (rule 3), and organisation name similarity (rule 4). For example, rule 1 matches organisation names records in a specific postal code area in The Netherlands for the city of Amsterdam.

Rule 1 specifies record pairs with postal codes that match exactly within a country and city. In this rule, the number of records an organizational name variant has in correlation with a specific postal code, is taken into account. This measure is important because an organization with only a few records assigned to a specific postal code area is suspected to be a false positive result. Another measure in this rule, is the percentage of records an organization name label has in a specific postal code, in relation to the total number of records associated with a certain organisation name. This percentage is also used to filter out organization name variants with low values on this measure.

Rule 2 is defined as record pairs with email domains that match exactly. The email domain labels should have a minimum count in the database. Records coupled by rule 3 share the same organisation type in a city. Organisation names that could not be typed in the pre-cleaning are excluded.

Rule 4 scores the level of string similarity of two organisation names within a city with the Levenshtein distance. The intuitive definition of Levenshtein distance is the amount of edits one needs to perform to change one organisation name into another organisation name. The implementation of Levenshtein distance described here uses the edit distance to calculate (in %) how similar is one string to another string. The use of the Levenshtein distance is based on the premise that organisational names that score above a certain threshold value, say 95% or so, can be considered as similar, and are therefore paired.

Rules 5-9 are combinations of rules 1-4. The combined rules have stricter thresholds than the basic rules, so the rules could match on different records.

2.2.2 Score Record Pairs (Step 2b)

Pairs of records are scored in step 2b. A record pair is described as two records that have scored on at least one rule and therefore share meta-data. Records can score on multiple rules, i.e. the scoring is additive. Therefore, the total score for record pairs has to be determined.

An example with 5 records, their active rules, and their total scores is presented in Figure 2. A line between two circles indicates a record pair. For example, the two records for ‘Vrije Univ Amsterdam’ and ‘VU Amsterdam’ share the same postal code, email domain, and organisation type, within the same country and city. Rule 4 does not fire, the string similarity is considered too low. Therefore, this record pair receives 4 points (see Table 1). The other records in the example are scored in the same way and are represented with a connecting line.

In step 2c, record pairs above the threshold value, total scores ≥ 4 in Figure 2, are included for the clustering algorithm. The threshold value is increased for geographical areas with a high number of organisations, to prevent the potential erroneous coupling of records pairs. The rules scores express the strength of a certain rule. Furthermore, the more rules that are active for a publication pair, the more evidence there is that two different organisation names are indeed variants of each other. In the example scoring system, only rules 8 and 9 are strong enough to solely pass the threshold value. However, for a pair of records often combinations of rules are required to exceed the threshold value., e.g. see the link between ‘Univ Amsterdam’ and ‘Univ Hosp Amsterdam’ in Figure 2.

2.2.3 Cluster Records (Step 2c)

Matched records pairs, i.e. record pairs with a score above the threshold, are clustered by means of single-linkage, hierarchical, clustering in step 2c. In Figure 2, for example, the records 'Univ Amsterdam' and 'Univ Hosp Amsterdam' are a matched pair, and the records 'Univ Amsterdam' and 'Emma Childrens Hosp' are a matched pair. The clustering algorithm makes a link between these two initial clusters via the joint record 'Univ Amsterdam', by merging the two clusters into a new cluster with three records, depicted by 'Cluster 2' in Figure 2, and so on. The final cluster will represent the (partial) history of name variants of an organisation. In the figure, there is not enough proof for the clustering of 'VU Amsterdam' with 'Univ Hosp Amsterdam', this is indicated by a dotted line. Therefore, two clusters are created by the algorithm, representing the two different universities in the city of Amsterdam. Notice that, if the threshold is increased, e.g. more clustering is induced, resulting in on average smaller cluster sizes.

2.3 Post-processing

In the post-processing stage, non-clustered records are labelled as separate clusters and added to the results to give a complete overview. Finally, tables are created that provide detailed summary information about the clusters. An example of such a table, which gives a cluster description, is given in Table 2. A combination of relational support tables, provide a good basis to work with the results of the clustering in practical data analysis.

3 CASE STUDY

In this case study, the method is used for the cleaning of scientific organizations present in the Web of Science (2015) bibliographic database. Bibliometric databases are large databases that are used to study the growth of scientific publications, patterns of collaboration, the impacts of science, and evidence-based performance assessment. For most of these analyses, it is necessary to increase the data quality by cleaning the relevant tables.

Cleaned organizational names are important for the Leiden Ranking (2015), produced by the Centre for Science and Technology Studies (CWTS, 2015). The CWTS Leiden Ranking 2015 offers insights into the scientific performance of 750 major universities worldwide, based on indexed research publications obtained from the Web of Science. This university name identification process is carried out manually and is therefore time-consuming and cumbersome. In the manual process, organizational labels are clustered and after that unified. This method can be trusted as very accurate because every organizational label that is under investigation, is verified with the help of the Internet and with other means, in order to be concluded as a name affiliation of a certain scientific organization. These cluster are a 'golden set', and used as a benchmark for the clusters produced by the automatic method in a precision-recall analysis.

In Table 2, the partial cluster for Leiden University in The Netherlands is depicted to show the end product of the clustering method. Each cluster is identified by a 'cluster_id' and is composed out of one or more records, that show supportive meta-data.

Table 2: Example cluster with id '3717' with name variants for 'Leiden University', ordered by the number of scientific publications, labelled by 'n_pubs'.

cluster_id	nu	ny	nc	nc_no	n_pubs	org_type
3717	NETHERLANDS	LEIDEN	LEIDEN UNIV	6739	69006	Univ
3717	NETHERLANDS	LEIDEN	LEIDEN STATE UNIV	3753	3701	Univ
3717	NETHERLANDS	LEIDEN	LEIDEN UNIV HOSP	20484	3480	Univ
3717	NETHERLANDS	LEIDEN	STATE UNIV LEIDEN	853	2919	Univ
3717	NETHERLANDS	LEIDEN	UNIV LEIDEN HOSP	42225	2231	Univ
3717	NETHERLANDS	LEIDEN	LUMC	550897	1505	Univ
3717	NETHERLANDS	LEIDEN	UNIV HOSP LEIDEN	7657	1251	Univ
3717	NETHERLANDS	LEIDEN	Leiden Univ Med Ctr	780691	1096	Univ
3717	NETHERLANDS	LEIDEN	UNIV MED CTR	177905	189	Univ
3717	NETHERLANDS	LEIDEN	UNIV MED CTR LEIDEN	216329	115	Univ
3717	NETHERLANDS	LEIDEN	Leids Univ	764733	95	Univ
3717	NETHERLANDS	LEIDEN	Leids Univ Med Ctr	1045872	49	Univ
3717	NETHERLANDS	LEIDEN	Leiden Univ Med Ctr LUMC	1740207	26	Univ

The precision and recall performance values for the best clusters per scientific organisation in the golden set are depicted in Figure 3, where the organisation names on the x-axis are ranked based on precision-recall values. The cluster with the highest value for the F1 measure, defined as the harmonic mean of precision and recall, is taken as the best cluster. In addition, the numbers in Table 3 show on average a precision of 0.95 and a recall of 0.80 for the best cluster.

Table 3: Average values of evaluation metrics for the best cluster in the Leiden ranking data set.

	Precision	Recall	F1
Best cluster (mean)	0.95	0.80	0.84
Best cluster (median)	1.00	0.89	0.98
Best 3 clusters (mean)	0.91	0.86	0.83

This shows that the clustering method is conservative, it chooses precision above recall. If the 3 best clusters for an organization are used in the evaluation the average recall is pushed to 0.86, with a slightly lower average precision. This indicated that for a number of

organisations the name variants are spread over a number of accurate clusters. Clusters with a lower precision are, in general, clusters belonging to very large cities, where multiple research institutes can be found in a relatively small area, which makes name normalisation more difficult.

4 CONCLUSIONS

In this research we have presented an efficient general rule-based scoring method for the clustering of name variants of organizations in large databases. The rules are based on organisation name similarity and meta data in the context of the organisation, like: country, postal code, email domains, organization type, etc. Basically, the method can work with any piece of relevant meta-data, as long as it is shared between records. Multiple rules can be combined to link organization names, because of the scoring system. The more rules that hold for a pair of organisation names, the more evidence there is that the organisation names are indeed valid name variants of each other. In other words, the rules in the system strengthen each other. Moreover, the rules are easy to understand and combine. Incorrect matching of organisation names is partly prevented by lowering the scores for certain sensitive rules and by increasing the threshold values, for example, for geographic locations with a high number of organisations.

Based on the results of the case study, it can be stated that the clustering method is careful, it values precision (on average 95%) over recall (on average 80%). In general, precision and recall are lower for areas with a high number of scientific organisations. Name variants of organizations might be split over multiple clusters, if there is not enough evidence for coupling names variants together. However, these alternative clusters do have a high precision and are therefore useful for analysis.

In conclusion, the method can be viewed as a general method for data cleaning, because it can be used to other types of data, e.g. person or author name disambiguation (Caron and Van Eck, 2014), as long as there is relevant meta data available. In future research, the cleaning method should be tested on multiple databases with name variants to find optimal values for scores and thresholds, and to improve the quality of the method for very large cities. In addition, we want to push recall performance forwards by further integrating string similarity measures (Cohen et al., 2003) in the method.

ACKNOWLEDGEMENTS

I thank Vasileios Stathias and Nees Jan van Eck for their contributions to this research. In this study I used the database facilities of the Centre for Science and Technology Studies (CWTS, 2015).

REFERENCES

Caron, E., van Eck, N.J., 2014. Large scale author name disambiguation using rules-based scoring and clustering, *In Proceedings of the 19th International Conference on Science and Technology Indicators*, pages 79-86, Leiden, The Netherlands.

Cohen, W., Ravikumar, P., & Fienberg, S., 2003. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, Vol. 3, pp. 73-78.

CWTS, 2015. *Centre for Science and Technology Studies*, http://www.cwts.nl, Leiden, The Netherlands.

De Bruin, R., Moed, H., 1990. The unification of addresses in scientific publications. *Informetrics*, 89/90, 65-78.

Koudas, Nick & Marathe, A. & Srivastava, D., 2004. Flexible string matching against large databases in practice. *Proceedings of the 30th VLDB Conference*.

Leiden Ranking, 2015. *CWTS Leiden Ranking 2015*, http://www.leidenranking.nl, The Netherlands.

Leiden University, 2015. http://www.leidenuniv.nl, Leiden, The Netherlands.

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

Maletic, J. I., & Marcus, A., 2010. Data cleansing: A prelude to knowledge discovery. In *Data Mining and Knowledge Discovery Handbook* (pp. 19-36). Springer.

Morillo, F., Santabárbara, I., & Aparicio, J., 2013. The automatic normalisation challenge: detailed addresses identification. *Scientometric*s, 95(3), 953-966.

Patstat, 2015, EPO Worldwide Patent Statistical Database, http://www.epo.org.

Song Y., Huang J., Councill I., Li J., & Giles C., 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '07). ACM, New York, NY, USA, 342-351.

Web of Science, 2015. *Thomson Reuters*, United States. http://www.webofscience.com.

APPENDIX

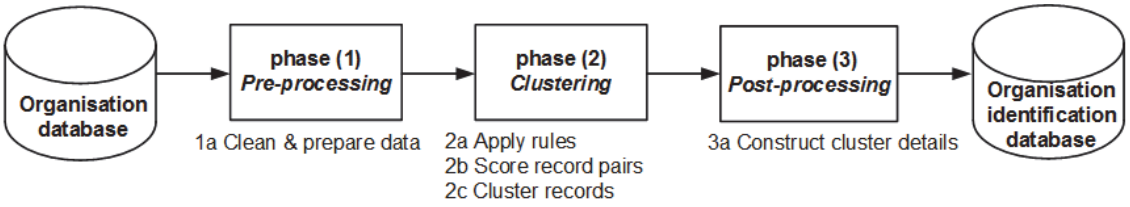

Figure 1: Stages in the identification process of organization name variants.

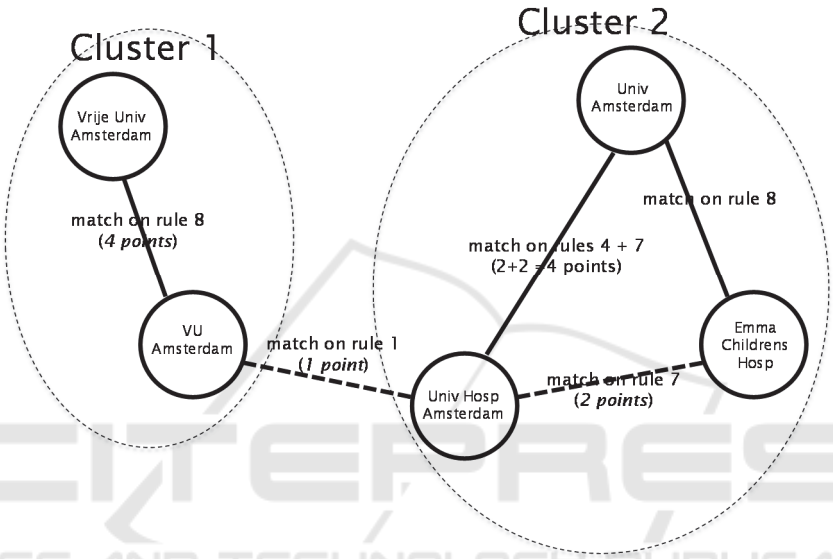

Figure 2: Scoring and clustering example for the city of Amsterdam (with threshold ≥ 4). Amsterdam has two universities the Vrije University Amsterdam (Cluster 1) and the University of Amsterdam (Cluster 2).

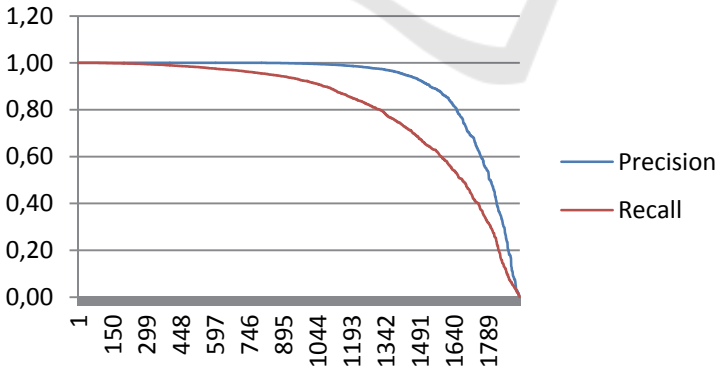

Figure 3: Precision (upper line) and recall (lower line) analysis on the Leiden Ranking data set.