Efficient supervised and semi-supervised approaches for affliations disambiguation

Pascal Cuxac, Valérie Bonvallot, Jean-Charles Lamirel

Efficient supervised and semi-supervised approaches for affiliations disambiguation

Pascal CUXAC[1], Jean-Charles LAMIREL[2] and Valerie BONVALLOT[1]

[1]pascal.cuxac@inist.fr ; valerie.bonvallot@inist.fr
INIST-CNRS, Vandoeuvre les Nancy, France

[2]jean-charles.lamirel@loria.fr
LORIA-Synalp, Vandoeuvre les Nancy, France

Abstract

The disambiguation of named entities is a challenge in many fields such as scientometrics, social networks, record linkage, citation analysis, semantic web...etc. The names ambiguities can arise from misspelling, typographical or OCR mistakes, abbreviations, omissions... So the search of names of persons or of organization is difficult, a single name can appear in different forms.
This paper proposes two approaches to disambiguate on the affiliations of authors of scientific papers in bibliographic databases: the first way, considers that we have a training corpus, and uses a Naive Bayesian model. The second way assumes that we have not resource learning, and uses a semi-supervised approach, mixing soft-clustering and Bayesian learning. The results are encouraging and are already partially applied in a scientific survey department. However, we aware that our approach may have limitations: we can't process efficiently highly unbalanced data but solutions are possible for future developments.

Introduction

In bibliographic databases, affiliations of authors are of paramount importance. Hence, they permit to the laboratories or institutes to get national and even international visibility, as well as they consequently provide authors with scientific caution. We cannot discuss the issue of affiliations without talking of "Shanghai ranking" which aims at evaluating universities. Our purpose here is not to feed up the controversy (Van Raan 2005) (Liu & Cheng and Liu 2005) , but to point out that the treatment of affiliation plays an important role in the calculation of universities "performance".
Moed (Moed 2005) reports some problems with author's names and also institutions: "*Authors from the same institution, or even from the same department, may not indicate their institutional affiliations in the same way*". Depending on the country, it is not always clear how to name a laboratory with respect to its supervisory authorities. The

affiliation is also important information to disambiguate author names in bibliographic databases. In this regard Wang notes: *as the amount of available information increases, problem caused by misspelling, spelling difference, and name or affiliation change also become worse.*" (Wang and al. 2012).
A standardization of data in bibliographic databases is necessary to carry out informetrics studies, but it is not a trivial task: the practice, intentional or otherwise, of omitting institutional affiliations, or giving incomplete or wrong information is not uncommon (Hood and Wilson 2003).
This paper proposes an approach based on Nave Bayes learning method and overlapping clustering. It is structured as follows: section 2 summarizes related works and identifies problems. Section 3 describes our approach firstly with supervised learning method and then with semi-supervised method. Next section 4 reports experiments and results. Section 5 conclude and discusses future work.

State of the art and discussion:

As part of bibliometric analyzes, the authors affiliations can produce statistics by laboratories as well as by institutes or universities. However, such analyses often face with problems of high variability and heterogeneity of naming: a single laboratory name may thus appear in several different ways if the authors use different abbreviations, incomplete or misspelled words (typing mistakes, spelling...). In addition, some universities may have several names (for example University Pierre and Marie Curie = University Paris VI). This problem is known for long but still persist nowadays. In the 1990s, De Bruin et al. (De Bruin and Moed 1990), point out the problem of variability of the author addresses in databases such as SCI (Science Citation Index). They highlight the case of countries like Germany or France where the heterogeneity of data is particularly important. Zitt (Zitt and Bassecoulard 2008) emphasizes the importance of data standardization (author names, affiliations) with special consideration to countries like France where affiliations overlap is important (for example one laboratory may have a University affiliation and CNRS affiliation). For many bibliometrics analysis, the unification of institutional addresses is an essential task, often boring, to be carried out prior to any study ((Bourke and Butler 1996), (Osareh and Wilson 2000)).
For solving the problem, De Bruin (De Bruin and Moed 1990) propose to deal separately with all the words belonging to affiliations and to use in a second step a classification strategy to unify all possible variations of the different words. In a later work (De Bruin and Moed 1993), the same authors use a "single-link clustering" approach to delineate different areas of science on the basis of affiliations. French et al. (French & Powell and Schulman 2000) supply an authority file after a cleaning step (name of country, zip codes, states, expansion of abbreviations, acronyms ...) and then use a clustering based on an "edit distance". Recent approaches also address the problem by the single use of NLP methods, like in (Galvez and Moya-Anegn 2006).
The terms data cleaning, data scrubbing, data standardization, data disambiguation, data homogenization and also entity resolution are used to refer to the tasks of transforming source data into clean or normalized data for loading in databases or linking with other data sets or computing statistical indexes (bibliometric analysis for example). If as we

have seen, these problems are essential in bibliometrics, they are also recurrent in many other areas where the heterogeneity of data is an important problem. This can be within a file or a database, but also when combining information from heterogeneous sources (e.g. record linkage). Erhard Rahm (Rahm and Do 2000) classifies data quality problems encountered in data cleaning tasks (fig. 1). In our case, we can assimilate the multi-source to a bibliographic database reporting papers from journals of different publishers.

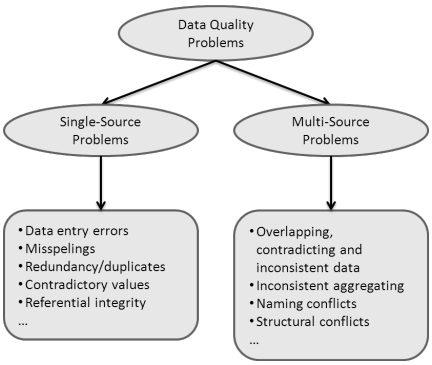

Figure 1: **Data quality in data cleaning tasks figure from (Rahm and Do 2000)**

The approach presented by Fellegi and Sunter (Fellegi and Sunter 1969) is the reference widely used in record linkage to identify the same entities in different datafiles. It is based on calculating similarity scores between two records. Generalizations of this method for more than two data files have been recently proposed (Sadinle & Hall and Fienberg 2010) (Sadinle and Frienberg 2012). Ventura (Ventura & Nugent and Fuchs 2012) mixes disambiguation and record linkage algorithms, using Random Forest, and applies this methodology to a case study of inventors of USPTO patents in the optoelectronics field. In this context of record linkage, Churches (Churches and al. 2002) shows that probabilistic Hidden Markov models for pre-processing data (names and address), give accurate results with complex data such as residential addresses.
When training data is available many of this studies use metrics to measure similarities between data such that Jaccard, soft-TF-IDF and mainly edit distance (Huang & Ertekin and Giles 2006). In his review paper, Bilenko compares the performance of various matching methods and concludes than learned affine edit distance can outperform others with EM techniques (Bilenko and al. 2003).
However, probabilistic approaches have been proposed by several authors like e.g. Carayol (Carayol and Cassi 2009), which proposes a bayesian approach to treat the who's who problem in European patents. He arises the transitivity issue that we'll discuss in the conclusion. Conversely, some authors propose approaches based only on unsupervised algorithms. This is the case of Niu (Niu & Wu and Shi 2006) which presents a new method for entity disambiguation using textual information and interobject relationship to evaluate similarity. Entities are author names, and interobject relationship is related to coauthorship network.
The novel methodology developed by Ashwani for unification of authors names use web mining to get full names and find publications pages (Aswani & Bontcheva and Cunningham 2006).
As we see, applications are numerous, be it in bibliographic data bases, in data ware-

houses, in multiple record linkage, in semantic web, or also in semantic digital libraries as shown in (Jiang and al. 2011). To conclude lets quote the standardization actions conducted through the International Standard Name Identifier (ISNI) and the Virtual International Authority File (VIAF). The aim of ISNI (http://www.isni.org) is to identify international public identities of individuals or communities and to provide tools for disambiguation. VIAF (http://viaf.org/) is a research project of OCLC (Online Computer Library Center) which aims to align lists of authorities (including proper names) to form an international reference database.
Do not forget to mention the Oyster software: is an open-source software developed by Talburt in the ERIQ Research Center (Entity Resolution and Information Quality at the University of Arkansas). OYSTER (Open sYSTem Entity Resolution) is an entities resolution system using XML scripts (Zhou and al. 2010).

Our approach:

We present two different methods for affiliation disambiguation: first, a supervised learning approach relying on manually analyzed reference dataset, and in a further step, a semi-supervised approach whose goal is to get rid of a training corpus for cases where no validation data are available.

Supervised learning method:

Supervised learning methods allow to produce rules from a learning corpus, generalizing what they could learn to the unknown inputs. In the literature there are many methods such as SVM, Rocchio, K-NN, Naive Bayes, HMM, Decision Trees...Our supervised approach is based on a Naive Bayes (NB) algorithm.
Let C a set of affiliations classes $C = \{c_1, c_2, ..., c_k\}$, the problem is to assign to an affiliation, one of these categories. Using a set of N labeled affiliations $\{(a_i, c_i), 1 \leq i \leq N\}$, we construct a classification function $\mathcal{F} : A \rightarrow C$ with A = set of all affiliations.

Bayes' formula for a given affiliation a allows to calculate its probability of belonging to a particular class c :

$$P(c \mid a) = \frac{P(a \mid c) * P(c)}{P(a)}$$

with
$P(c \mid a)$ = probability of c given a,
$P(a \mid c)$ = probability of a given c,
$P(a)\ P(c)$ = respectively probability of a and probability of c

If we simplify by assuming that labels are randomly distributed (are not dependent on the length of the affiliation or the position within the affiliation), then the probability of affiliation a given a class c, is

$$P(a \mid c) = \prod P(w_i \mid c)$$

with w_i = the i-th word of a.

then, by applying the Bayes rule we can classify an affiliation in a class c:

$$c = arg\,max\ P(a \mid c)P(c) = arg\,max\ P(c \mid a)$$

Despite of the two main known defects of such method, that are, its ignorance of the order of the words and its ground hypothesis that words are independent conditionally to their class membership, its represents a good alternative for solving our problem. The results obtained by this implementation of Bayes theorem are valid and demonstrated by (Hand and Yu 2001). Hence, Domingos et al (Domingos and Pazzani 1996) formerly showed that the misclassification error of NB is minimized as compared to other methods.

Method for semi-supervised classification:

Whenever no a priori knowledge is available, we applied a semi-supervised methodology. In this case, we firstly process by the use of an overlapping clustering method. The exploited clustering technique is the axial k-means algorithm (a variant of the k-means method proposed by Lelu, (Lelu 1993), which allows to produce clusters presenting particular characteristics:

- they can overlap because the clustering method allows an object or a variable to belong to more than one cluster;
- the constituting elements of a cluster, objects and variables, are ranked by decreasing similarity with the cluster ideal type.

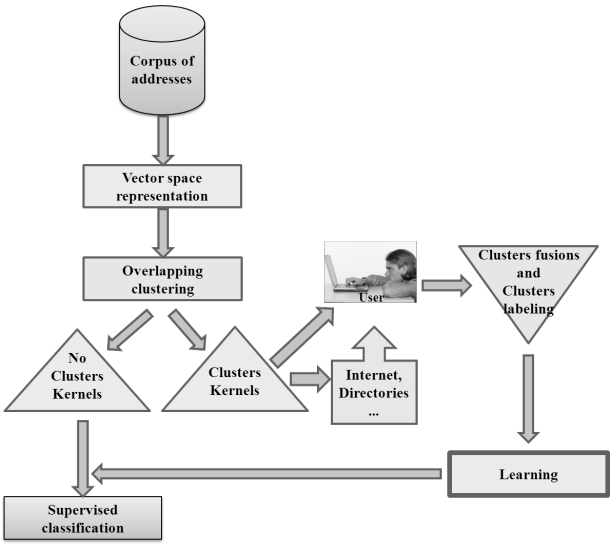

Figure 2: **Semi-supervised classification schema.**

In the second step, we only retain the major representatives of the classes, which are the documents that have the highest values of projection on the axes representing the classes.

These documents will then be used as the training corpus. Then, we calculate the most representative words of each class and we use each of these groups to query the web (via Google). The site that represents the most relevant answer of the search engine for a given class is used to label the class. If necessary, we proceed to a further step of classes merging.
In the last phase, we train the NB method with the corpus defined in the clustering phase and labeled by the class names extracted from the web. The testing process is achieved on the complementary corpus of documents eliminated after the clustering phase.
Figure 2 summarizes in a schema, our semi-supervised approach with the three steps: soft-clustering, clusters labelling and clusters fusion, and finally the Naive Bayes classification.

Experiments

In this section we present our results obtained with three corpora and the two methods presented previously.

Datasets:

We used three different datasets: a first dataset of 10 057 French affiliations (noted here-after A1), a second small dataset of 150 Lorraine affiliations (region of France) (noted A2), and a last dataset of 2266 French affiliations extracted from WOS and SCI (noted A3). All those datasets have been preprocessed by splitting the affiliations into words using space as separator (any punctuation, including dash, is firstly removed). Given the difficulty we had to cut affiliations into words in the datasets A1 and A2 (dash, sometimes missing space ...), we have then used a second splitting technique based on n-grams after converting affiliations in string without spaces or punctuation. Supervised learning is applied on dataset A1, and semi-supervised one is applied on the two other datasets.

LOR1A INRIA CNRS UMR 7503 BP 239	LORIA
LORIA INRIA Lorraine 615 rue du Jardin Botanique	LORIA
LORIA UHP Campus scientifique BP 239	LORIA
LAB-PLANETOL-GRENOBLE, GRENOBLE 9	IPAG
CNRS, UJF, OBSERV GRENOBLE, ASTROPHYS LAB, F-38400 ST-MARTIN-DHERES	IPAG
UNIV-GRENOBLE-1, CNRS, LAB ASTROPHYS GRENOBLE LAOG, UMR 5571, GRENOBLE	IPAG
OBSERV-GRENOBLE, F-38041 GRENOBLE	IPAG
UNIV-GRENOBLE-1, LAB ASTROPHYS GRENOBLE, INSU CNRS, GRENOBLE	IPAG
LAB-ASTROPHYS-GRENOBLE, GRENOBLE	IPAG
LAB-ASTROPHYS-OBSERV-GRENOBLE, GRENOBLE	IPAG
UNIV-STRASBOURG, INST PLURIDISCIPLINAIRE HUBERT CURIEN, CNRS, IN2P3, STRASBOURG	IPHC
UNIV-STRASBOURG, IPHC, CNRS, IN2P3, STRASBOURG	IPHC
ULP, IPHC, IN2P3, F-67037 STRASBOURG	IPHC

Figure 3: **Data sample with various forms of addresses**

Figure 3 illustrates a data sample with the address in the first column and the laboratory acronym in the second column. We can see three laboratories presented in different way.

Measures of performance:

The results are evaluated in terms of recall, precision, F-measure, because we know a priori classes of all affiliations.

$$Recall : R = \frac{TP}{(TP + FN)} \quad Precision : P = \frac{TP}{(TP + FP)} \quad F - measure : F = \frac{2 * P * R}{(P + R)}$$

where TP, FP, FN, mean number of true-positives, number of false-positives, and number of false-negatives, respectively.

Supervised learning:

The dataset A1 was split into training dataset and test dataset successively represented by the words of affiliations and by the n-grams. The figure 4 shows the distribution of the affiliations in the 53 resulting classes (test + train) and thus highlights the fact that the resulting classification is highly unbalanced.

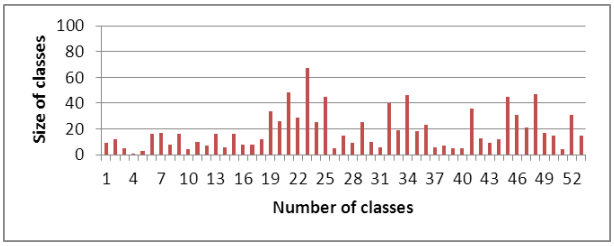

Figure 4: **Distribution of affiliations in resulting classes (dataset A1)**

The classification results on the dataset A1 are presented in Table 1.

Table 1: **R, P and F values for corpus A1.**

	Recall	*Precision*	*F-measure*
N-gram	0.92	0.94	0.93
Words	0.81	0.88	0.85

Because of the problem of individualization of words affiliations in dataset A1, the results obtained with n-grams appear slightly better, with an optimum of Recall of 0.92. It should also be noted that a systematic lookup of the analysis results of NB which got a very high probability, whilst being in contradiction with the expected (i.e. human labeled) results, permits us to prove that the manual labeling of the test corpus was sometimes wrong (the model was giving the right answer in all that cases!).

Semi-supervised classification:

The datasets A2 and A3 were used for this experiment. The said datasets are split into train and test with the methodology described in the former section detailing our approach. The figure 5 reports the distribution of affiliations in the resulting classes (A2: 19 classes; A3: 10 classes). It highlights that the smaller dataset (A2) is highly unbalanced, whilst the bigger one (A3) is homogeneous.

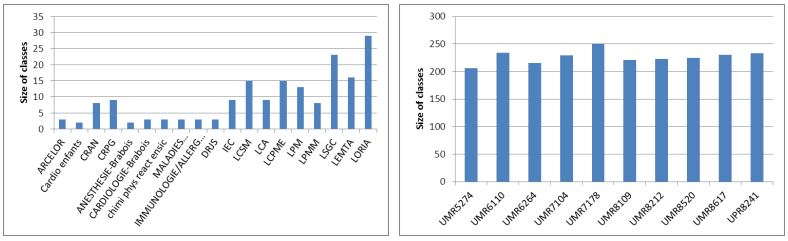

Figure 5: **Distribution of affiliation in resulting classes (datasets A2 and A3)**

This distribution is obviously intentional, in order to be aware the impact of data distribution on our results.

Table 2: **: Results of K-means R, P and F values for corpora A2 and A3.**

Kma only	*Recall*	*Precision*	*F-measure*
Corpus A2	0.44	0.81	0.55
Corpus A3	0.40	0.95	0.54

With this kind of data, the results of K-means are not performing (table 2). It is probably due to the bad representation of data. As we discuss in the last section of this paper, a vector space representation taking in account the scientific content linked to each addresses should improve the clustering result.

Table 3: **: R, P and F values for corpora A2 and A3.**

	Recall	*Precision*	*F-measure*
Corpus A2	0.79	0.76	0.73
Corpus A3	0.98	0.97	0.97

As detailed in table 3, the results are very good for the dataset A3, but significantly lower for the corpus A2, where recall is average. This is due to the large number of classes as compared to the small size of the corpus and to the strong imbalance of these classes. Indeed, in this case the exploited clustering method is becoming blind to small classes.

Conclusions and discussions

The results we obtained with our approach for affiliations disambiguation are very encouraging, both in the supervised learning context and in the semi-supervised one. Our experiments also permit us to show that our method provide a significant assistance for correcting the results of human labeled affiliations. However, it is clear that we must practice more experiments to conclude on an overall relevance of the methodology. Hence, there still remain some weaknesses in our methodology, mainly related to the exploited clustering method in the case of very unbalanced classes. We thus plan to conduct tests with other clustering methods and implement data balancing techniques. Another important point would be to exploit a learning method that should be able to learn with

character strings of variable length. It would be necessary to develop an automatic method that could highlight the conflicting cases. Future works should also take into account the xml structure to consider separately the cities, street names, the laboratories names...
Another way is to consider the scientific content of documents, such as titles and abstracts of articles published. Once these documents indexed each address would be represented by a vector of words (describing research activities) allowing probably a more relevant classification.
The study of transitivity can perhaps permit to detect false positive or false negative results and thereby isolate the results to be verified. We also propose to compare our results with those obtained using the Oyster software.

References

Aswani, N. & Bontcheva, K. and Cunningham, H. (2006). Mining Information for Instance Unification. *Lecture Notes in Computer Science*,4273, 329-34.

Bilenko, M. & Mooney, R. & Cohen, W. & Ravikumar, P. and Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5), 16-23.

Bourke, P. & Butler, L. (1996). Standards issues in a national bibliometric database: The Australian case. *Scientometrics*, 35(2), 199-207.

Carayol, N. & Cassi, L. (2009). Whos Who in Patents. A Bayesian approach. http://hal-paris1.archives-ouvertes.fr/hal-00631750

Churches, T. & Christen, P. & Lim, K. and Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, 2, DOI:10.1186/1472-6947-2-9

De Bruin, R. E. and Moed, H. F. (1990). The unification of addresses in scientific publications. *Informetrics* 1989/90, 6578. Elsevier Science Publishers, Amsterdam.

De Bruin, R. E. and Moed, H. F. (1993). Delimitation of scientific subfields using cog nitive words from corporate addresses in scientific publications. *Scientometric*s, 26(1), 65-80.

Domingos, P. and Pazzani, M. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *International Conference on Machine Learning (ICML)*, 1996, 105112.

Fellegi, I. and Sunter, A. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

French, J.C. & Powell, A.L. and Schulman, E. (2000). Using Clustering Strategies for Creating Authority Files. *Journal of the American Society for Information Science and Technology*, 51, 774-786.

Galvez, C. and Moya-Anegn, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics*, 69(2), 323-345.

Hand, D. J. and Yu, K. (2001). Idiots BayesNot So Stupid After All? *International Statistical Review*, 69(3), 385-398.

Hood, W. and Wilson, C. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58(3), 587-608.

Huang, J. & Ertekin, S. and Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. *PKDD'06. LNAI*, 4213, 536544, Springer-Verlag.

Jiang, Y. & Zheng, H.-T. & Wang, X. & Lu, B. and Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology*, 62(6), 1029-1041.

Lelu, A. (1993). *Modles neuronaux pour l'analyse de donnes documentaires et textuelles.* PhD, University Paris 6.

Liu, N. C. & Cheng, Y. and Liu, L. (2005). Academic ranking of world universities using scientometrics:A comment to the Fatal Attraction. *Scientometrics*, 64(1), 101-112.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation.* Springer.

Niu, L. & Wu, J. and Shi, Y. (2012). Entity Disambiguation with Textual and Connection Information. *Procedia Computer Science*, 9(0), 1249-1255.

Osareh, F. and Wilson, C. S. (2000). A comparison of Iranian scientific publications in the SCI: 1985-1989 and 1990-1994. *Scientometrics*, 48(3), 427-442.

Rahm, E. and Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4),3-13.

Sadinle, M. & Hall, R. and Fienberg, S. (2010). Approaches to Multiple Record Linkage. *cscmuedu, http://www.cs.cmu.edu/˜rjhall/ISIpaperfinal.pdf*

Sadinle, M. and Fienberg, S. E. (2012). A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record-Systems. arXiv:1205.3217. http://arxiv.org/abs/1205.3217

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133-143.

Ventura, S. L. & Nugent, R. and Fuchs, E. R. H. (2012). Methods Matter: Revamping Inventor Disambiguation Algorithms with Classification Models and Labeled Inventor Records. *SSRN eLibrary.* http://papers.ssrn.com/sol3/papers.cfm?abstract id=2079330.

Wang, J. & Berzins, K. & Hicks, D. & Melkers, J. & Xiao, F. and Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, 1-21.

Zhou, Y. & Talburt, J. R. & Su, Y. and Yin, L. (2010). OYSTER: A Tool for Entity Resolution in Health Information Exchange. *Proceedings of the 5th International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO 2010 E-BOOK)*, 358-364.

Zitt, M. and Bassecoulard, E. (2008). Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics*, 8, 49-60.