Author disambiguation

Brian Ampwera
Universite Lumiere Lyon 2

Sandra Pietrowska
Universite Lumiere Lyon 2

***Abstract*—Different authors may share the same names, either as full names or as initials and last names. Our goal is to perform the author disambiguation that enables accurate assignment of publications to a certain author profile. In this paper we describe the algorithm considering the fullnames, disciplines and institutions. We perform two experiments in order to verify the commutativeness of our approach and the profile extraction of the authors validated by the manually disambiguated dataset.**

I. Introduction

The name disambiguation finds its application in many areas such as evaluating faculty publications, calculating statistics of social network and author impacts as well as citation analysis. However, it poses a challenge because of several reasons: while the names and their translation may have different alternative structures and meanings. Using the metadata enables to construct the profiles of distinct authors. Manual processing is costly and time-consuming especially for the large-scale databases, therefore we propose an automatic algorithm in order to solve the disambiguation problem. The aim of this paper is to present our contribution on determining the identity of the individuals enabling the effective paper assignment. The main difficulty of the task is to define the possible variety of the names format, meaningful selection of attributes (metadata) as well as setting the threshold that does not discard the correct classification. These obstacles are to be overcome by using the integrative framework presented in the following sections. Furthermore, considering the institution variable, the difficulty is that the authors institutions are not assigned distinctively and can only be derived as probabilistic near approximations which is discussed. The algorithm presented in the paper covers 4 major steps and is performed on the preprocessed data. First of all we define the clusters of names that resemble to each other. Then we perform the split on disciplines, followed by clustering by institutions. The last step is deriving the full names that constitute the final assignments. This article is organized as follows: section 2 describes the data used in the research and the preparation methods. Section 3 explains in details the steps of algorithm including the mathematical aspects. Section 4 presents the results and the evaluation method. The last section consists of two parts: a conclusion and recommended further work.

II. Data preparation

The processed dataset comes from Web of Science and the collection period pertains 4 years: from 2007 to 2010 inclusive. It contains 318591 unique publications, 2328539 authors, 261741 institutions and 250 fields presented in tables (attributes given in brackets): articles (id_paper, authors), articles_authors (id_paper, author number, authors name), institution (id_paper, nb of institution, institution name given in several tuples). Since the set is very large, the first step is to analyze the given dataset in order to have an overview of the problem and detect the complexity of the task. The descriptive statistics of the data is presented in the table I. The maximum number of publications per author name is 458 (name is Banerjee S), while the average number of publications by author is approx. 5. It highlights the treated problem of disambiguation, hence acting with the knowledge of the real-world distribution of papers and authors within 4 years (data collection period) we make a logical assumption that there are more individuals that have this name. The number of outliers is 56361 that means that they significantly distant from the first or third quartile value. We face the same problem with the number of authors publications per year. The descriptive statistics shows that there are maximum 145 publications assigned to one authors name while 75% of the ranked set of data by author name is 1. Taking into consideration the above results, one may draw conclusion that there is a high authors name ambiguity. Moreover, the queries performed on the tables show that there are 17 publications without the authors given, so viewing the low fraction, though the significance, we have excluded them from the dataset. Basing on the above results, we proceed data preparation on the following variables: name (surname, first letter of a given name and middle name) field and institution. First of all, from the table articles_authors we have selected the authors names. In order to derive the surname we reduce the inflected name composed from first name, surname and middle name to their stem that is the surname. The second variable, field, contains 250 distinct attributes and no preparation was needed. Thirdly, we have considered the given institution names that were already stemmed. There were 261741 distinct entries of institutions. Nevertheless, to ensure that the dataset contains the distinct attributes we have removed the stopwords, that are in this case meaningless and produce noise, such as: ctr, dept, fac, lab, sch, univ. Moreover, viewing that some names included the same words, in further steps, we would proceed with computing the distance between them. As a result, we are interested in finding the possible declination of the name and the assignment of the individuals to the paper so the entity relationship diagram is presented on Fig.1.

Att	Min	Max	Mean	1st quartile	Median	3rd quartile	Outliers
Publications per author name	1	458	4.87	1	2	4	56361
Authors publications per year	0	145	1.22	0	0	1	205167
Average publications per author in a year	0.25	114.5	1.22	0.25	0.5	1	56361

TABLE I
DESCRIPTIVE STATISTICS

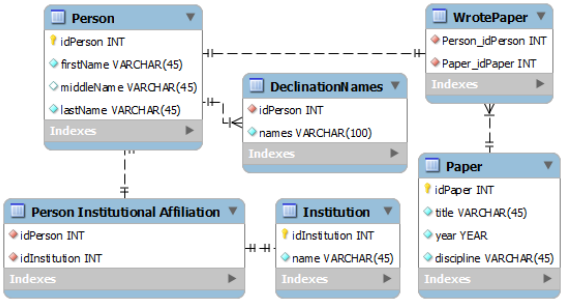

Fig. 1. Entity relationship diagram.

III. METHOD

The disambiguation algorithm consists of 4 steps presented consecutively in this section. The major idea that we have followed is to apply algorithms sequentially as a decision tree.

1) Create the n clusters Si with the papers written by the authors with the same surname
2) Create the m clusters Di from the cluster Si , where i = 1,2...n with the papers belonging to the same discipline
3) Create the j clusters Ii from the cluster Di where i = 1,2...m with the papers published in the same institutions
4) For each cluster Ii Derive the full names and assign a probability that relates the paper, surname and name The method is presented using the running example.

A. Split by author

The root node of a decision tree contains all the authors surnames. It is worth mentioning that since the disambiguation does not find its use within the names that appear only once we discard these values from the dataset. Then, we form the clusters Si, (where i= distinct surname) using the following algorithm: for each name N which is a set of unigrams of the form X1,X2...Xn we remove Xn and convert the name N-Xn to lower case obtaining the clusters Si each containing the papers published by the same author (surname). This partition allows us to create 147378 none singular sets of clusters. For example from the forms Daumas A, Daumas M, Daumas S, Daumas L we obtain the format daumas or from de Murcia JC we obtain the format de murcia. Using the running example, by this step, the result contains the set S(daumas) consisting of 14 appearances of surname daumas in the dataset. The motivation for this is that if there is a typographical error in the unigram Xn , this method would account for it.

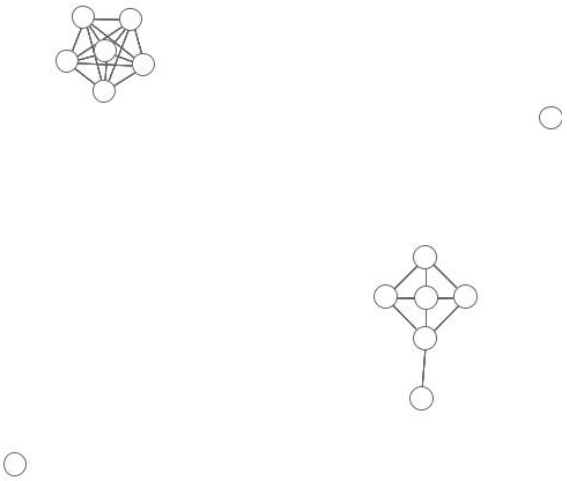

Fig. 2. Connected components for discipline split.

B. Split by discipline

The aim of this step is to divide each of the previously computed n sets into the subsets Di with respect to disciplines. All the articles has an attribute discipline and may contain more than one discipline. Following up on the idea of an interactive global science map using the ISI Subject Categories (Leydesdorff and Rafols, 2009; Rafols et al., 2010), our objective in this step is to generate an equivalently interactive map for the disciplines on each surnames cluster Si. Based on the distance between the disciplines we create the new subsets Di using this algorithm: for each paper P in Si we compute the discipline distance with respect to P2,P3...Pv and as a result obtain the matrix M(v x v). Then, since on the diagonal the distance is always 1 we transform it into an adjacency matrix of papers of size (n2 - n)/2. We obtain the set of clusters visualized as an undirected graph (Graph 1) with the sets of connected components.

The clusters have been created using a threshold computed with a following assumption: during an authors lifetime the discipline distance between his papers will lie within 2 standard deviations from the mean distance of said authors papers. Each set is the resulting cluster. For the running example we have obtained 4 clusters presented on Fig.2.

Knowing that the same author may write the papers that have different field origins, these fields are likely to be assigned to the same clusters given that the distance between the fields is small enough. The conclusion drawn from this is

Name groups	People count	Relative Frequency	Rcf
100	231777	0.985454808	0.985455
200	2114	0.008988172	0.994443
300	800	0.003401389	0.997844
400	179	0.000761061	0.998605
500	112	0.000476195	0.999082
600	73	0.000310377	0.999392
700	38	0.000161566	0.999554
800	15	6.37761E-05	0.999617
900	18	7.65313E-05	0.999694
1000	16	6.80278E-05	0.999762
more	56	0.000238097	1

TABLE II
FREQUENCY DISTRIBUTION FOR AUTHOR NAME GROUPS

Paper Author Count	Frequency	RF	CRF
10	194774	0.82812779	0.82812779
50	32856	0.139695065	0.967822856
100	4179	0.017768008	0.985590864
200	2107	0.00895841	0.994549273
300	800	0.003401389	0.997950663
400	182	0.000773816	0.998724479
500	110	0.000467691	0.99919217
600	58	0.000246601	0.999438771
700	31	0.000131804	0.999570575
800	18	7.65313E-05	0.999647106
900	19	8.0783E-05	0.999727889
1000	19	8.0783E-05	0.999808672
1100	10	4.25174E-05	0.999851189
More	35	0.000148811	1
Total	235198		

TABLE III
FREQUENCY DISTRIBUTION FOR PAPER AUTHOR NAME GROUPS

Cluster	Paper id	Discipline
1	139094	nvironmental Sciences, Geosciences Multidisciplinary.
2	261383	Neurosciences,Psychology, Experimental
	106385	Behavioral Sciences,Neurosciences
	111832	Orthopedics,Surgery
	177081	Behavioral Sciences,Neurosciences
	190056	Geriatrics & Gerontology,Gerontology
	32111	Clinical Neurology,Neurosciences
3	270924	Infectious Diseases,Microbiology
4	260631	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic
	294249	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic
	213826	COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS,Physics, Mathematical
	73766	Computer Science, Software Engineering,Mathematics, Applied
	147519	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic
	307003	Computer Science, Artificial Intelligence

TABLE IV
SPLIT BY DISCIPLINES

that given that the 2 clusters contain the disciplines not related to the others (given the threshold) we define these authors as 2 different individuals. The first is the author that has written the paper with id 139094 which is Daumas S and id 270924 that is Daumas A. On the 2 remaining clusters we perform the step 3, that is the split by institution.

C. Split by institutions

This step involves the division of the previously computed clusters Di into subsets according to the assigned institution in order to obtain the clusters Ii . Each paper has w institution names, however they are not assigned to concrete authors. Given a set of institutions for each paper we create the subsets Ii containing the papers that are described with the same authors name, discipline and institution respectively. We apply this method to obtain the set of clusters Ii. Since the same author may publish the paper in different institutions, being aware of it we solve this problem by making probabilistic estimation based on that the institutes that have the highest frequency in a set that emerges. The most frequent institution

is the basis of the split. The illustration of this step on the running example is presented in the Table V.

D. Name generation

Once we have the final clusters Ii that contain the papers from the same field and institutions we derive the full name of the author. We call it the possible declinations of the full names and regarding the number of different formats we assign the corresponding probabilities Probablitiy of Author = Frequency of a declination/Total Number of Declinations that are stored in the database as presented on Fig.3.

IV. Experiments

A. Experiment 1

The goal of presented algorithm is to successfully disambiguate the authors. The statistics presented in section 1, shown that there were numerous authors surnames with large number of papers. But not only, these authors have to be disambiguate. There may be the a group of individuals within the set of few papers with the same surnames. The experiment 1 aims at showing how the steps given in section 3 deal with the given problem. It aims at performing the experiment on the randomly selected papers for a given name and obtain a decision YES/NO if the paper is written by the author. It attempts to demonstrate the commutative property of the algorithm. We have chosen 1 random author profile, we select 10 random papers from the dataset and basing on the algorithm decide either it is his paper or not. We repeated the experiment 100 times respectively. This is repeated for 100 times (hence, 100 authors, for each 10 papers, 100 times). The experiment execution workflow is as follows:

1) consider the institution of the paper if yes accept else no
2) consider the discipline distance if the discipline distance below the threshold of 0.3
3) consider the surname if the surname is the same.
4) Consider the declination probability for the name, if lower assign NO else assign YES

We repeated this experiments using an incremental set of random author profiles and the obtained results are presented in Table VI.

Number of authors selected	Accuracy
100	0.982
500	0.978

TABLE VI
Experiment 1: results

B. Experiment 2

The second experiment was based on the manually disambiguated data. Firstly, we took the name id of the paper provided in the dataset. The results are presented in Table VII.

From the first experiment the suggested heuristic shows that if the algorithm disambiguates the data set then the results are significant, however when applied to the entire data set for large clusters of surnames the performance is very low. This can be attributed to the large amount of noise in the institution set. We assume that a cleaner set of institutions would make the inferences of the author better.

V. Results and evaluation

The algorithm use the three main information given in the dataset which are the name, discipline and institution. We have performed two experiments. In order to verify if the algorithm is consistent we have changed the order of the operands. It resulted in obtaining the accuracy for the sets 100 and 500 - 0.982 and 0.978 respectively. The conclusion is that it attempts to be commutative. The second experiment, performed on the manually disambiguated data has shown that for the small sets the results have a high accuracy, while the bigger sets produce noise and effectively, the worse results. After performing the algorithm on the entire dataset we have obtained the following results:

In order to further the research, we have designed a user interface. This system allows to browse through the papers allowing to find authors with the given title of a paper. The interface has been designed with thought of possible improvements of the algorithm, namely allowing the algorithm to learn and as a result improve the classification.

VI. Conclusion

In this article we have investigated the algorithm for authors disambiguation. We have analyzed the dataset by computing the descriptive statistics in order to have an overview on the task. Then we have prepared the data for further processing, that includes the basic text mining operations such as removing the stop words and stemming. Then we had performed an algorithm consisting of 4 major steps, carried out sequentially with respect to discipline and the institution that gave us the separate clusters of papers with the same authors name assigned. The last step involved defining the possible declinations with the corresponding probability. The algorithm has been evaluated with two experiments that showed good performance on low size sets. In conclusion the algorithm is sufficient for small data sets but proves computationally expensive for larger data sets. Although the initial dataset contained more variables such as keywords, address, references and journals, we proved that given the split on the three chosen variables (name, field and institution respectively) has an satisfiable accuracy for small sets. Therefore, the initial assumptions are plausible. Moreover the given split preserves the dataset from overfitting, where the model describes random error or noise instead of underlying relationships.

A. Further work

The further exploration may be made on several aspects. First of all the authors disambiguation does not involve the problem of misspelling or the common change of the surname. Secondly, since this is unsupervised learning, it is difficult to perform an absolute, meaningful evaluation. On that note,

Cl	Paper id	Discipline	Institution
2	261383	Neurosciences,Psychology, Experimental	serv chirurg orthoped & traumatol=2
	106385	Behavioral Sciences,Neurosciences	
	111832	Orthopedics,Surgery	
	177081	Behavioral Sciences,Neurosciences	anim umr 5169 toulouse 3=2
	190056	Geriatrics & Gerontology,Gerontology	
	32111	Clinical Neurology,Neurosciences	
4	260631	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic	eliaus perpignan
	294249	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic	
	213826	COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS,Physics, Mathem	
	73766	Computer Science, Software Engineering,Mathematics, Applied	
	147519	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic	
	307003	Computer Science, Artificial Intelligence	

TABLE V
SPLIT BY INSTITUTIONS

Cluster	Discipline	Institutions	Names(Pr)	pers_id
D_3	Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS,Physics, Mathematical Computer Science, Software Engineering,Mathematics, Applied Computer Science, Hardware & Architecture ,Engineering, Electrical & Electronic Computer Science, Artificial Intelligence	cnrs lirmm um 2-2, math texas arlington-2, eliaus perpignan-6	Daumas M (1.0) Daumas M(1.0) Daumas M(1.0) Daumas M(1.0) Daumas M(1.0) Daumas M(1.0	260631 294249 213826 73766 147519 307003
D_0	Neurosciences,Psychology, Experimental Behavioral Sciences,Neurosciences Orthopedics,Surgery Behavioral Sciences,Neurosciences Geriatrics & Gerontology,Gerontology Clinical Neurology,Neurosciences	chu amiens nord serv chirurg orthoped & traumatol-2, cnrs rech cognit anim umr 5169 toulouse 3-2	Daumas S (1.0) Daumas S (1.0) Daumas S (1.0) Daumas L (0.33) Daumas A (0.33) Daumas S (33)	261383, 106385, 111832 177081, 190056, 32111
D_1	Environmental Sciences,Geosciences, Multidisciplinary		Daumas S	139094
D_2	Infectious Diseases,Microbiology		Daumas A	270924

Fig. 3. Split by names.

Cluster	Test Set	Full Set	Discipline Clusters	Institutions	Declinates	Accuracy
abbott	15	202	7	12	17	43.22
boussaid	19	19	4	5	7	100
daumas	14	14	4	4	4	87.5
daumas-duport	19	19	1	2	2	65.1
el din	4	4	1	1	1	100
jezequel	134	134	23	42	82	83.1
muller	34	1689	[NA]	[NA]	[NA]	[NA]
velcin	4	4	1	1	2	100
zighed	11	11	1	1	2	100
Grand Total	257					

TABLE VII
RESULTS FOR TESTS USING MANUAL DISAMBIGUATION

it would be recommendable to build a validation dataset that spans various fields and institutions where the authors institutions are concretely defined. Thirdly, even though the assumptions about the split on a distance are plausible, a community detection algorithm would provide better results. Finally, the future work may be directed into learning algorithm that would create the links of the most commonly visited papers given the authors.

REFERENCES

[1] Kai-Hsiang Yang, Hsin-Tsung Peng, Jian-Yi Jiang, Hahn-Ming Lee, Jan-Ming Ho *Author Name Disambiguation for Citations Using Topic and Web Correlation*, 2008

[2] L. Leydesdorff, I. Rafols *Interactive overlays: A new method for generating global journal maps from Web-of-Science data*, 2012, Journal of Informetrics.

[3] William E. Winkler, *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*, 1990

[4] Vladimir I. Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, 1966