Author Disambiguation Using Multi-Aspect Similarity Indicators

Thomas Gurney[1], Edwin Horlings[2] and Peter Van Den Besselaar[3]

[1] *t.gurney@ratheanu.nl (Corr. Author),* [2] *e.horlings@ratheanu.nl*
Rathenau Instituut, Anna van Saksenlaan 51, 2593HW, The Hague (The Netherlands)

[3] *p.a.a.vanden.besselaar@vu.nl*
VU University Amsterdam, De Boelelaan 1105, 1081HV, Amsterdam (The Netherlands)

Abstract

The use of scientometrics is becoming increasingly prevalent in many forms of scientific analysis and policy making. Key to accurate bibliometric analyses is the ability to correctly link individuals to their corpus of work, with an optimal balance between precision and recall when querying the larger dataset in which their corpus resides. We have developed an algorithm that addresses the issues of discarded records due to null data fields, and have implemented a dynamic approach to similarity calculations based on all available data fields. We have conditionally filtered the data to account for differences in author contribution and age difference between publications, resulting in a higher recall and precision of returned records. We present results from a test dataset of heterogeneous catalysis publications. Preliminary results demonstrate high F-scores and improvements on stand-alone techniques.

Introduction

The use of scientometrics has become increasingly prevalent in many forms of scientific analysis and policy making. Key to good bibliometric analysis is the ability to correctly link individuals to their respective corpus of work, with an optimal balance between precision and recall when querying the larger dataset in which their corpus resides. The most common problem encountered is that of multiple persons having the same last name and initial. Other problems include misspelled names, name abbreviations and name variants. Within a small dataset, these errors can be corrected using manual checks. However, with large datasets time and labor constraints severely hamper disambiguation efforts. The increasing scale and scope of scientometric studies and the rapid rise of Asian science systems (Phelan 1999; Moed, Glänzel et al. 2004; Trajtenberg, Shiff et al. 2006; Cassiman, Glenisson et al. 2007) – where variance in names is substantially lower – reinforce the need for an automated approach to author disambiguation.

There is a need for algorithms designed to extract patterns of similarity from different variables, patterns that can set one author apart from his or her namesake, and link to other data sources. Our primary focus in this paper is the problem of correctly identifying multiple persons sharing the same last name and initial. We have developed a novel algorithm that increases the precision and recall of author specific records, whilst decreasing the number of records discarded due to null indicators. The algorithm takes into account factors such as author contribution, time difference between publications and dynamic combinations of indicators used.

Realistic expectations of disambiguation techniques

Techniques for author disambiguation are based on the assumption that the source data (whilst not providing a unique identifier for every author) at least maintain a correct spelling of the last and first name. This assumption has been proven to be naïve in almost all data repositories, as there are multiple avenues for error to creep in. As a result we have chosen to focus on one spelling of the last name only.

Literature Review

The current literature on disambiguation is split between computer science and sociological and linguistic approaches. There have been few papers melding the two approaches. They are discussed briefly in this section.

Zhu et al. (2009) constructed string- and term-similarity graphs between authors based on the publication titles. Graph-based similarity and random walk models were applied with reasonable success to data from DBLP. A similar study by Tan et al. (2006) uses search engine result co-occurrence for author disambiguation. Yang et al. (2008) discovered disambiguation problems in citations and developed a method to determine correct author citation names using topic similarity and web correlation with the latter providing stronger disambiguation power. In disambiguating researcher names in patents, Raffo and Luillery (2009) investigate the different search heuristics and devise sequential filters to increase the effectiveness of their disambiguation algorithm.

Perhaps the most relevant recent publications come from Tang and Walsh (2010) and Onodera et al. (2011). Using the concept of cognitive maps and approximate structural equivalence, Tang and Walsh developed an algorithm based on the knowledge homogeneity characteristics of authors. They analysed the effectiveness of their technique on two common names (one of English origin, the other of Chinese origin). Their technique was remarkably successful, but biased in that records that did not exhibit any similarities in the cited references were treated as isolates and thus excluded from the results. Onodera et al.'s study is the most similar to ours in that they use similarity probabilistic techniques. But they differ in their objective of disambiguation as they aim to retrieve specific authors' documents, not to discriminate between different authors within a dataset. They report a high success rate of 85% in retrieving the correct publications per author. Unfortunately, Onodera et al. pre-filtered possible retrieved records by eliminating those that did not possess the correct contingent of metadata. As a result their success rate was biased.

It should be further noted that, generally speaking, most of the studies mentioned here rely on relatively few personal and relational data fields. It is all too common in studies that records are excluded from further analyses if they do not possess the full complement of metadata. This is unfortunate as the discard rate is rather high in some cases (close to 50% in the case of Tang and Walsh (2010)). The issue of discarded data due to lacking indicators is rarely, if ever, discussed.

Method

Method overview

The objective is to create a network of publication/author nodes in which edge strengths are the probabilistic value of the two nodes being the same person, as calculated by logistic regression. A community detection algorithm is employed over the network to further discriminate the pairings of nodes.

Data

In the testing and implementation phases of this project we have used data related to heterogeneous catalysis collected by a project team within the PRIME ERA Dynamics project. The dataset is a collection of 4979 articles, letters, notes and reviews featuring 5616 authors. The records were retrieved from ISI's Web of Science and parsed using SAINT (Somers *et al* 2009). Through manual cleaning and checking, each publication was assigned to the correct author. Each record is considered unique, and is based on a combination of the article and author IDs which were assigned during the parsing process. There are 3872 different last names and of these there are 2014 last names which have more than one

publication. There are 4403 author last name and first initial variants, with 208 instances in which more than one author has the same last name and first initial. We have focused our efforts on the instances in which there are more than one author with the same last name.

Metadata fields used and similarity calculations

In regard to the discarding of data, we have chosen to include all records, no matter their complement of metadata. A dynamic approach is required where, prior to the similarity calculations, a listing of shared meta-data is created, including the age difference between publications (also in Onodera et al. (2010) and the individual author contributions of each publication.

The age difference between publications will have an effect on the degree of similarity between publications as there may be a change in the individual's research focus over time. Author contributions are calculated using the sum of the fractional author counts of the author positions of the two records where the contribution of the second and last authors is equal to 2/3 of the contribution of the first author. Any other authors contribute 1/3 of the first author. This is normalised so that the sum of all the fractions is equal to 1. For example, in a publication of 6 people, where a is the contribution of the first author:

$$a + 2/3a + 1/3a + 1/3a + 1/3a + 2/3a = 1; \text{ and } a = 3/10 \text{ (Moed 2000).}$$

In the case of alphabetical listings of authors, each author is assigned a value of $1/n$ (where n is the total number of authors). The addition of author contribution as a filter follows the reasoning that an author's position within the publication may influence the selection of title words, cited references, keywords, journal choice and more.

The shared available metadata for each pair of calculations is referred to as the Null Combination (NC) code and each pair of unique author/article comparison calculations are based on this code. The year differences (YD) and average author contributions (AAC) are categorised (YD - 7 categories; AAC - 4 categories). The similarity calculations are based on the Tanimoto coefficient – τ – and follows the form $\tau = N_{AB} / (N_A + N_B - N_{AB})$, where N_A is the count of tokens in A, N_B is the count of tokens in B and N_{AB} is the count of tokens shared between A and B.

The metadata fields for which the Tanimoto coefficient is calculated are: title words, abstract words, last names and initials of coauthors, cited references in string form, normalized keywords (author and indexer assigned), normalized research addresses and journal categories.

Logistic Regression

The initial similarity calculations are based on comparing records that have the same first letter of their last name. Use the first letter introduces more possibilities for matches between different authors as this is a relatively small dataset. Computationally speaking, this approach is acceptable, but for larger datasets (+10 000) the comparisons would be based on matching last names. Logistic regression requires the presence of two pre-determined groups. Having previously identified the authors' correct publications, we were able to create an input dataset in which the pre-determined groups are defined as Group 2 (identical authors) and Group 0 (non-identical) as shown in Table 1.

Table 1. Input data table for regression analysis.

			Independent variables				*Filters*		
Group	*A*	*B*	*-1*	*-2*	*-3*	*(...)*	*YD Category*	*AAC Category*	*NC Code*
0	1	5	a	b	c	...	x	y	123
0	1	6	NULL	b	c	...	x	y	23
2	2	3	NULL	b	c	...	x	y	23
2	1	3	NULL	b	NULL	...	x	y	2

An initial regression analysis was performed using only the NC as a filter, i.e., only rows of records with specific NC codes were included, and from this the optimum combination of available independent variables was determined through comparing the -2 Log Likelihood values of each NC code variation and choosing the variation that returned the best results.
The data was split into calibration and testing sets in an approximately 75:25 ratio to test the validity of the model. Once the validity was confirmed all available data was used in further analyses. A second regression was run with the newly selected optimum NC codes, and the year difference and average author categories as filters. The full regression formula is as shown in Equation (1).

$$\ln(\mathbf{Y}/1\text{-}\mathbf{Y}) = \beta 0 + \beta_1(\text{SimCoauth}) + \beta_2(\text{SimAbstract}) + \beta_3(\text{SimTitle}) + \beta_4(\text{SimCitedRef}) + \beta_5(\text{SimAuthorKeywords}) + \beta_6(\text{SimIndexerKeywords}) + \beta_7(\text{SimRes.Address}) + \beta_8(\text{SimJournalCategory}) \textbf{\textit{ under condition }} \textit{(NC;YD;AAC)} \qquad \textit{(Equation 1)}$$

The coefficients found in the final regression are used to estimate the pair probabilities of the data set. The flowchart in Figure 1 summarises the order of operations in which the calculations are performed.

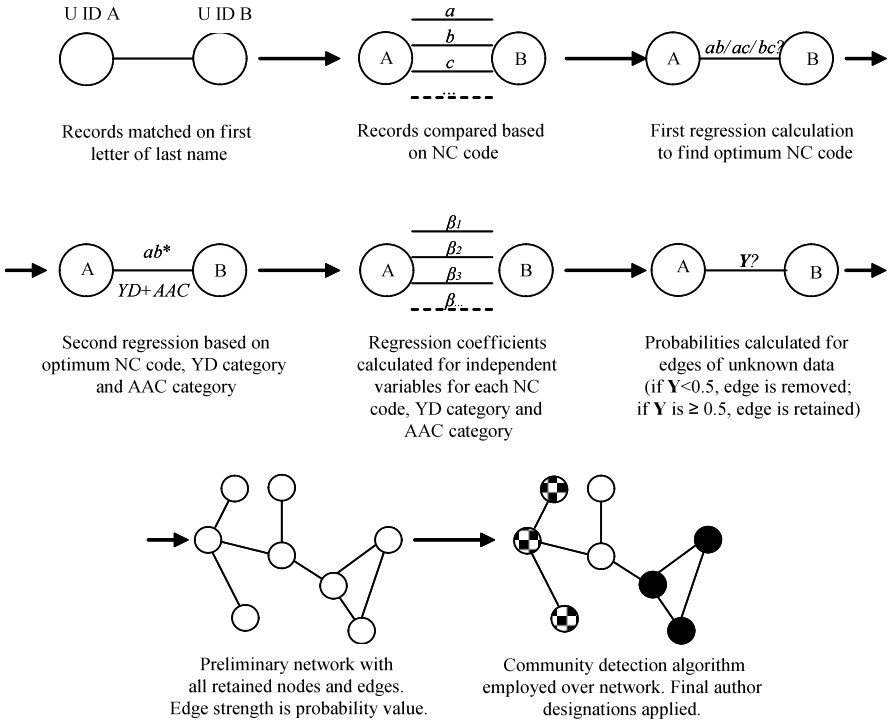

Figure 1. Summary of order of operations of data processing, regression calculations and final author disambiguation

Final Author Assignment

Final author designation is performed by the community detection algorithm of Blondel et al. (2008). This algorithm takes into account the weighted edges of a network and assigns each node to a specific community based on the surrounding nodes and their edge weights. Logistic regression predicts the probability as to whether two publications are from the same author on a row by row basis, but the community detection algorithm works on the entire interconnected network of nodes or publications.

Results

For our results we have chosen names of different origin and varying publication counts to demonstrate the effectiveness of the algorithm. As seen in Table 2, using only the logistic regression technique results in perfect recall for all the publications, but exhibit low precision rates. Using the community detection algorithm, we achieve very high precision rates with resulting F-measures ranging from 0.86 to 1.00 as seen in Table 3.

Individual Name	*LI,C*	*LI,BT*	*LI,JL*	*LI,W*	*LI,XB*	*GARCIA,H*	*GARCIA,J*	*GARCIA,R*
# of Original Publications	15	12	10	7	6	107	3	4
# of Retrieved Publications	64	16	64	64	64	116	116	116
True Positives (Tp)	15	12	10	7	6	107	3	4
False Positives (Fp)	49	4	54	57	58	9	113	112
False Negatives (Fn)	0	0	0	0	0	0	0	0
Precision	0.23	0.75	0.16	0.11	0.09	0.92	0.03	0.03
Recall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F-Measure	**0.38**	**0.86**	**0.27**	**0.20**	**0.17**	**0.96**	**0.05**	**0.07**

Table 2. Disambiguation results for two author last names using logistic regression only.

The low precision rates using only logistic regression highlight the fact that different authors within a sub-network may be highly clustered around each other, but with one or two very weak (0.5-0.6 probability score) edges linking the clusters. Thus, as seen in the Garcia examples, precision is very low with authors who have few publications in a large sub-network. However, with the application of the community detection algorithm, these weak edges are taken into account and the author clusters are separated out, resulting in very high precision rates.

Individual Name	*LI, C*	*LI, BT*	*LI,JL*	*LI,W*	*LI,XB*	*GARCIA,H*	*GARCIA,J*	*GARCIA,R*
# of Original Publications	15	12	10	7	6	107	3	4
# of Found Publications	15	12	10	7	6	109	3	3
True Positives (Tp)	15	12	10	7	6	107	3	3
False Positives (Fp)	0	0	0	0	0	2	0	0
False Negatives (Fn)	0	0	0	0	0	0	0	1
Precision	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75
F-Measure	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**0.86**

Table 3. Disambiguation results for two author last names using logistic regression and community detection algorithm.

Discussion and conclusions

The results in this study demonstrate accounting for time differences between publications and the retention of all possible metadata as independent variables. More important is that an author's contribution to each publication ultimately affects what title words, abstract words, and cited references for example, are used. This is a very important factor when considering similarity-based disambiguation methods such as ours.

A drawback of this method surfaces when individuals publish in multiple, unrelated fields. Unless there are bridging publications, that exhibit similarities to more than one distinct publishing field, the networking aspect will show separate clusters, thus affecting recall.

With the benefit of further research, we will investigate the minimum number of publications necessary to consistently recreate these results and develop a trans-disciplinary calibration set.

A benefit of the algorithm is the scalability of data where the only limit to size is the relational database software limits. This algorithm and technique could be applied further to most forms

of entity resolution, such as that of inventors and applicants in the patenting field. We hope to develop it in such a form soon.
Author ambiguity is a serious enough issue to warrant more attention. We hope that through our method we will be able to improve upon past efforts and to eventually present a user-friendly, open-source tool for scientists, policy-makers and evaluators, so that decisions based on error prone results become less common. We aim to integrate this disambiguation tool into SAINT (available from reference website). This would allow records from various data repositories to be parsed and accurately sorted by author/inventor on the order of hundreds of thousands of records.

References

Cassiman, B., P. Glenisson, et al. (2007). "Measuring industry-science links through inventor-author relations: A profiling methodology." Scientometrics **70**(2): 379-391.
Moed, H. F. (2000). "Bibliometric indicators reflect publication and management strategies." Scientometrics **47**(2): 323-346.
Moed, H. F., W. Glänzel, et al., Eds. (2004). Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems Dord-recht, Kluwer Academic Publishers.
Onodera, N., M. Iwasawa, et al. (2011). "A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search." Journal of the American Society for Information Science and Technology **62**(4): 23.
Phelan, T. J. (1999). "A compendium of issues for citation analysis." Scientometrics **45**(1): 117-136.
Raffo, J. and S. Lhuillery (2009). "How to play the "names game": Patent retrieval comapring different heuristics." Research Policy **38**(10): 1617-1627.
Somers, A., T. Gurney, et al. (2009). Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics. The Hague, Rathenau Insitute. Download available from www.rathenau.nl.
Tan, Y. F., M. Y. Kan, et al. (2006). Search engine driven author disambiguation. 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, ACM.
Tang, L. and J. P. Walsh (2010). "Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps." Scientometrics **84**(3): 763-784.
Trajtenberg, M., G. Shiff, et al. (2006). "The Names Game: Harnessing Inventors' Patent Data for Economic Research." NBER working paper.
Yang, K. H., H. T. Peng, et al. (2008). "Author Name Disambiguation for Citations Using Topic and Web Correlation." Research and Advanced Technology for Digital Libraries: 185-196.
Zhu, J., X. Zhou, et al. (2009). "A Term-Based Driven Clustering Approach for Name Disambiguation." Advances in Data and Web Management: 320-331.