

Author disambiguation for enhanced science-2.0 services

Jeffrey Demaine

Institut für Forschungsinformation und Qualitätssicherung (iFQ) Bonn, Germany

The disambiguation of authors with homonymous names has been a growing concern in bibliometrics for several years. The identification of authors is a necessary precondition for accurate bibliometric calculations. The basic technique of matching authors' names based on ratios of shared letters is susceptible to error in the case of very common names or in broad research fields. A more advanced approach to the problem relies on the contextual metadata relating to an author. Using the co-author and the co-citation relationships that pertain to a given author, a map of the intellectual landscape of that author can be constructed. Such patterns of collaboration and citation are typically unique to each individual and serve as fingerprints for the identification of those who happen to share the same name. The automated analysis of contextual metadata for each author-name-instance can be used to enhance the management of knowledge within science-focused social media. This paper presents a minimalistic solution to the author disambiguation problem that leverages bibliographic metadata to generate clusters of articles belonging to distinct individuals who all share the same name. Implications for enhanced Science 2.0 services are discussed.

Keywords: *Author disambiguation; metadata, Science 2.0*

1. Introduction

The disambiguation of authors with identical (homonymous) names has been a growing concern in bibliometrics for several years. The identification of authors is a necessary precondition for accurate bibliometric calculations.

A review of the current literature on author disambiguation reveals three broad approaches to tackling the problem. The simplest technique involves matching the authors' names based on ratios of shared characters. [Bolacker *et al.* (1998); Lawrence *et al.* (1999); Galvez & Moya-Anegón (2007)] Some of the algorithms used include *Jaro-Winkler proximity*, the *Dice coefficient*, and *Levenstein distance*. While these techniques can go a long way in standardizing variations in author names, they are of little use in disambiguating correctly spelled but common names or in broad categories of science in which many researchers are active.

A more advanced approach to the problem relies on the contextual data relating to an author. [Tang & Walsh (2010); Kang *et al.* (2009); Wooding *et al.* (2006)] By exploiting the co-author and/or the co-citation relationships that pertain to a given author, articles can be grouped together so as to distinguish one author from another. Such patterns of collaboration and citation are typically unique to each individual and serve as distinguishing characteristics for the identification of those who happen to share the same name. Indeed, in such contextual approaches, the name of the author is not even taken into consideration by the disambiguation algorithm. This approach is therefore more robust to instances of spelling errors.

Beyond these lies the realm of statistical approaches to author disambiguation. [Zhang *et al.* (2007); Tang *et al.* (2008)] These offer the promise of a more accurate solution to the problem. In addition, one would expect that a statistical approach would offer some measure of probability as to the accuracy of a particular author-disambiguation solution, although the literature does not address this issue. While this third category offers the most

potential for highly accurate name disambiguation, the research at this level does not seem to be generalizable outside of small amounts of curated test data. These techniques often rely on *a priori* assumptions about the frequency of certain parameters or require a "training set" of correctly-identified authors before they can begin. For example, Smallheiser and Torvik (2009) estimate the statistical properties of the names being disambiguated as a precursor to the calculation of author-name groups by statistical methods. Soler (2007) presents a statistical model for disambiguation in which many assumptions and approximations are made as to the frequency of words and names. Such exercises achieve remarkable results with and may point the way to better disambiguation principles, but it is hard to see how these can be made to work without considerable fine-tuning or in-depth knowledge about the characteristics of the field in which the name to be disambiguated is found. If one must determine how many "J. Smiths" are working in the field of materials science in order to be able to disambiguate instances of "S. Johnson", one has simply exacerbated the problem.

A good example of a more robust solution is given by the recent work of Gurney, Horlings, and Van Den Besselaar [Gurney *et al.* (2011)]. Their approach incorporates up to eight different metadata elements and performs regression analysis to determine the strength of the similarity between each pair of metadata. A network of relationships is then constructed with the regression coefficients as the weight of the edges between items. A clustering algorithm then groups the network into author instances. Gurney *et al.* make the important point that many other author-disambiguation algorithms are evaluated against datasets that have been pre-filtered. The algorithms naturally perform very well because any records that do not contain all the requisite metadata are removed beforehand. In contrast, their approach makes do with whatever metadata is available. This is an important issue, as the solutions proposed for author-disambiguation are purely academic if they do not handle the imperfections of real-world data.

The algorithm presented here falls into the second category of author-disambiguation techniques in that it relies on metadata relating to an author's publications. This approach is decidedly simpler than many of the other techniques in the literature. It was developed with an eye towards providing a data-cleaning procedure that facilitates the work of the *Institut für Forschungsinformation und Qualitätssicherung* (iFQ). As part of a German research consortium the iFQ performs bibliometric studies in support of Germany's national science policy. To accomplish this, the iFQ has access to a database of the raw data that is behind the Web of Science. To ensure that the analyses are as accurate as possible, the iFQ requires a tool that helps to clean the data by disambiguating authors such that the correct articles can be attributed to the people and institutions being studied. However, all of the metadata for authors is not always available (depending on the publisher from which the data was collected). While some records contain complete information including the author's first name and institutional affiliation other records do not. Thus for the application described herein to work with all records, it must rely only on the lowest common denominator available: co-authors and references.

Thus the specific challenge addressed here is to group author-name instances in a reliable and automated manner – given a limited amount of metadata – such that it becomes practical to undertake large-scale studies involving many authors from any field of research. In this real-world scenario, utility is as important as accuracy.

2. Method

Homonymous name instances (retrieved from the iFQ's databases) that share co-authors and/or references are grouped together as one individual according to the following algorithm. This is implemented in Java and uses local caching of names and ID numbers to reduce the need to re-query the database for the same information.

1. Query based on a given last name and first initial ("SMITH J").
2. Create an array of the authorID numbers ("A"). Create a copy of this array ("B").
3. For each element in A (that has not been compared previously), create an "author-group" array, and then loop through all subsequent elements in B:
 - 3.1. For each combination of A and B, retrieve the article each author ("a1" and "b2") is associated with. Query for the co-authors of both articles, as well as for the references contained in each article a1 and b2.
 - 3.2. Create two arrays, one for each set of co-authors (a1_c and b2_c), as well as another two arrays, one for each set of references (a1_r and b2_r).
 - 3.2.1. Loop through the co-author arrays, comparing the names. Loop through the references, comparing the referenceIDs.
4. If, after comparing all co-authors and all referenced articles, a match has been found (at least two matches in the case of references), then add the authorID number of the author from B to the "author-group" belonging to author A.

3. Results

To test the accuracy of this application, data from a study of known German researchers was used to verify matches. For each researcher a list of their known publications was compared against the groups generated by the algorithm. The results were analyzed using a confusion table approach. The group with the largest number of matching publications was considered to be the predicted group for that individual (a "true positive"), and all other groups were therefore considered to represent other individuals (representing collectively the "negative" prediction). If a known researcher's paper appeared in one of the "other" groups, this is a "false negative" (it should not be in another group, but in the predicted group). Note that it was only possible to determine the Precision of the matches, as the total number of individuals with that name (and their publications) is not known.

	Condition (reality) TRUE	Condition (reality) FALSE	
Test (prediction) TRUE	True positive	False positive	Precision $TP \div (TP + FP)$
Test (prediction) FALSE	False negative	True negative	
	Recall $TP \div (TP + FN)$		

Table 1. A confusion table for evaluating the accuracy of matches.

Table 2 shows the names tested, the year from which those names were considered. For “B Borasoy” the algorithm created two groups. The first group held 17 of the articles known to have been written by this particular “B Borasoy”, and the second group contained ten more articles known to have been written by this very same person. Ideally all 27 articles would have been in the first group, resulting in a Precision of only 63%.

Name tested	Starting Year	Records retrieved	Groups identified	True Positive	False Negative	Precision
B Borasoy	1996	72	2	17	10	63.0%
R Huber	2002	798	64	45	2	95.7%
A Blaukat	1996	54	7	39	29	57.4%
G Behre	1995	180	5	30	13	69.8%
M Albrecht	1999	914	71	52	1	98.1%
L Ackermann	1999	111	13	32	10	76.2%
R Everaers	1994	69	4	16	5	76.2%
P Neumann	1997	454	76	50	0	100.0%
K Niebuhr	1996	18	7	4	4	50.0%
C Ochsenfeld	1996	65	4	20	7	74.1%
JW Pan	2000	257	11	47	9	83.9%
T Pietschmann	2002	99	2	28	2	93.3%
B Regenber	2002	25	1	20	0	100.0%
K Wiegand	1999	46	12	10	5	66.7%

Table 2. Accuracy of disambiguation over 14 names corresponding to a known individual of that name.

The average Precision for these analyses was 79%. However, the variability is quite high. One fairly common error seen in the results occurs when the same individual publishes articles with different sets of co-authors on topics that are disparate enough to not share any references. In such cases of “split personality” [Ley, 2009] the algorithm described above will leave two instances of the same person unmatched. This is a *false negative* in that the name-instance is assigned to another group when it is clearly the same person as a previously-created group to which it should have been added. One may decide to join these instances based on some other type of metadata, but without further supporting evidence, two homonymous names remain different individuals.

A more informative analysis of the results is possible when the exact number of individuals is known. For example, *Estel Cardellach* and *Esteve Cardellach* are both active researchers living in Barcelona. Estel works mostly in remote sensing while Esteve concentrates on Geology. The application identified 41 articles published by “Cardellach E” since 2002 (inclusive) and disambiguated them into four groups. If we take the largest group matching Esteve to be the positive prediction (group #1) and the largest group matching Estel is the negative prediction (group #2), then a confusion table can be constructed and the Precision and Recall of the disambiguation can be calculated¹.

¹ This type of table is often used in clinical studies to evaluate the effectiveness of some new drug (for example). The normal terminology is then “Condition” for the columns and “Treatment” for the rows.

Prediction	Reality		
	Esteve	Estel	
Esteve	25	1	Precision = 96%
Estel	3	12	
	Recall = 89%		

Table 3. Confusion table of disambiguation of “Cardellach E”

An evaluation of these results is given by the *Matthews correlation coefficient* of 0.79 [Matthews (1975)]. Interestingly, the single error in prediction (a *false positive*) occurred when Estel published an article on Geology, presumably citing or co-authoring with one of Esteve’s usual collaborators. The three *false negatives* were contained in a third group of exclusively Esteve’s articles. While these should have been collected into the first group with the other 25 articles by Esteve, the fact that they share no co-authors or references reveals a distinct research topic. Although not the goal of author disambiguation, these false negatives can be of interest in characterising the different aspects of a researcher’s career. Perhaps the individual is active in two distinct fields, or the separate groupings might indicate different phases of their career.

In comparison, a search of Thomson Reuter’s Web of Science online database retrieves 37 articles written since 2002 (inclusive) by “Cardellach, E”. Then using the Web of Science’s “Distinct Author Sets” feature, these 37 articles were grouped into 32 sets. This represents a simplification of only 13%. In addition, as we know that there are only two people with this name, the Recall is only 47% (Estel and Esteve each get one group, meaning that the 30 other groups are unnecessary).

Overall, the advantages of the approach presented here are two-fold: First, the application achieves fairly good results, especially considering that it only has two types of metadata to work with. Secondly, the application runs in a “hands off” manner, requiring only two pieces of information (a name and the year from which to begin collecting records) and a few minutes to organize the records into groups. This means that the application can be used by those who have no background in bibliometrics or programming. On the other hand, the limitations of this approach include the great deal of variability in the precision of the results and the inflexibility of the algorithm to account for publication patterns that do not fit the model.

As new versions of the database become available, additional types of metadata can be included into the algorithm to increase its accuracy. By extending the metadata used for comparison to include institutional affiliation it is likely that the reduction in superfluous names can be further improved. Because the spelling of the name to be disambiguated is not taken into consideration by the algorithm, the results are – theoretically – as good for very common names as for rarer ones.

4. Discussion

The minimalistic approach to author disambiguation presented here demonstrates how fairly accurate results can be obtained using only a limited amount of metadata. The technique for doing this should be of interest for the “Science 2.0” type of social media services such as *CiteULike* or *Mendeley* that are forums for sharing references and for building connections within the research community. These do not have all of the

bibliographic data that a database of publications has (such as the *Web of Science*), yet they still seek to provide a space where published research is organized.

Currently, CiteULike makes no attempt to disambiguate author names. One can browse CiteULike by author name, following the link from the author of one article to see a list of all articles by authors with that name. But each name is treated as a distinct entity, even if they refer to the same person. Using the system described here would allow a 2.0 service (such as CiteULike) to automatically create groups of author identities.

Indeed, linking information to create something greater than the sum of its parts is the whole point behind Science 2.0: by facilitating interactions within a given community, the conduct of research is accelerated and enhanced. In such an environment, a library or a virtual lab-space will have little control over the content. On the one hand this is a good thing: getting away from the top-down management of information and letting the research community communicate as it wishes. But at the same time, a certain amount of quality control is of benefit to all. Automated features that bring some order to the Science 2.0 environment are needed, particularly if they leverage the structure of the community or the communication. While author-disambiguation is not specifically a Science 2.0 problem, it provides an example of how uncertainty about identity (and about the research belonging to that identity) can be reduced algorithmically.

As CiteULike does not list the references belonging to each article, the part of the matching algorithm does not apply. However, one can turn the situation on its head and think of the users as the aggregating factor. If a user's profile lists several dozen articles as being of interest, then this is a type of citation. The names to be disambiguated could then be matched based on the profiles within which they are listed. When two articles written by a "Smith J" are listed in the same CiteULike user-profile, they are likely to be the same J. Smith.

If services such as CiteULike are to ever move beyond unsorted collections of documents, some structure must be automatically derived from their content. What is needed is an automated approach to author disambiguation such that as records are added to a bibliographic database, they are identified as belonging to the correct individual. Even if they were not 100% correct (in that a single "J. Smith" was split across two or more identities), the groupings would reveal the characteristics of the underlying data. This structure could then be exploited to provide an alerting or recommender service that enables users to follow an author of interest, notifying them when that author has published a new article.

5. Conclusion

The authority control of instances of author names is an ongoing concern for the accurate measurement of scientific publication patterns. Research into author-name disambiguation is pursuing a range of approaches ranging from matching on the syntax of the names to more techniques that employ advanced statistical analysis. Yet the latter is of no use in identifying individuals who share the same name, and the latter has only been applied to small, curated datasets (i.e. containing only records with the full complement of metadata). The minimalist approach presented here offers a balance between accuracy and practicality. It is fast and scalable. In addition, because it relies on only two types of metadata, it can be adapted for use in Science 2.0 web services such as CiteULike where users share uncurated information. The application described here suggests that it does not always have to be the

users that do the linking. Some automation (or a hybrid approach) would leverage the input of the community.

6. References:

Bollacker, Kurt D., Lawrence, Steve, Giles, C. Lee. (1998). CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. *2nd International ACM Conference on Autonomous Agents*. 116- 123.

Galvez, Carmen; Moya-Anegón, Félix (2007) "Approximate Personal Name-Matching Through Finite-State Graphs". *JASIST* **58**(13):1-17.

Thomas Gurney, Edwin Horlings & Peter van den Besselaar (2011). "Author disambiguation using multi-aspect similarity indicators." In: *Proceedings of ISSI 2011: Durban, South Africa. July 4-7, 2011*.

Kang, In-Su; Na, Seung-Hoon; Lee, Seungwoo; Jung, Hanmin; Kim, Pyung; Sung, Won-Kyung; Lee, Jong-Hyeok (2009) "On co-authorship for author disambiguation". *Information Processing and Management* **45**:84-97.

Lawrence, Steve; Giles, C. Lee; Bollacker, Kurt D. (1999) "Autonomous Citation Matching". *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, Washington, May 1-5, ACM Press, New York, NY, 1999.

Ley, Michael. (2009). DBLP – Some Lessons Learned. *Proceedings of the VLDB Endowment* **2**(2). <http://www.vldb.org/pvldb/2/vldb09-98.pdf>

Matthews, B.W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochim. Biophys. Acta* **405**, 442–451.

Smallheiser, Neil R.; Torvik, Vetle I. (2009) "Author Name Disambiguation". *Annual Review of Information Science and Technology (ARIST)*(B. Cronin, Ed.) **43**

Soler, José M. (2007) "Separating the articles of authors with the same name". *Scientometrics*. **72**(2):281-290.

Tang, Li; Walsh, John P. (2010) "Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps". *Scientometrics* **84**:763-784.

Tang, Jie; Zhang, Jing; Zhang, Duo; and Li, Juanzi. (2008). "A unified framework for name disambiguation." In *Proceeding of the 17th international conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 1205-1206. DOI=10.1145/1367497.1367728

Wooding, Steven; Wilcox-Jay, Kate; Lewison, Grant; Grant, Jonathon. (2006) "Co-author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis." *Scientometrics* **66**(1):11-21.

Zhang, Duo; Tang, Jie; Li, Juanzi; and Wang, Kehong. (2007). "A constraint-based probabilistic framework for name disambiguation." In *Proc. sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*. ACM, New York, NY, USA, 1019-1022. DOI=10.1145/1321440.1321600 <http://doi.acm.org/10.1145/1321440.1321600>