Scientometrics (2012) 91:435–449
DOI 10.1007/s11192-011-0589-1

Author disambiguation using multi-aspect similarity indicators

Thomas Gurney · Edwin Horlings · Peter van den Besselaar

Received: 5 December 2011 / Published online: 30 December 2011

Abstract 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. We have developed an algorithm that does this disambiguation task with a very high recall and precision. The method addresses the issues of discarded records due to null data fields and their resultant effect on recall, precision and F-measure results. We have implemented a dynamic approach to similarity calculations based on all available data fields. We have also included differences in author contribution and age difference between publications, both of which have meaningful effects on overall similarity measurements, resulting in significantly higher recall and precision of returned records. The results are presented from a test dataset of heterogeneous catalysis publications. Results demonstrate significantly high average F-measure scores and substantial improvements on previous and stand-alone techniques.

Keywords Author disambiguation · Precision and recall · Homonyms · Community detection · Data discarding

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

An extended version of a paper presented at the 13th International Conference on Scientometrics and Informetrics, Durban (South Africa), 4–7 July 2011 (T. Gurney, E. Horlings, P. van den Besselaar, 2011).

T. Gurney (✉) · E. Horlings
Rathenau Institute, Anna van Saksenlaan 51, 2593 HW The Hague, The Netherlands
e-mail: t.gurney@rathenau.nl

E. Horlings
e-mail: e.horlings@rathenau.nl

P. van den Besselaar
VU University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
e-mail: p.a.a.vanden.besselaar@vu.nl

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. This is especially important where bibliometrics is used for evaluation purposes. 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 labour constraints severely hamper disambiguation efforts. The increasing scale and scope of scientometric studies and the rapid rise of Asian science systems (Trajtenberg et al. 2006; Cassiman et al. 2007; Phelan 1999; Moed et al. 2004)—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 (first) 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 missing data. The algorithm takes into account factors such as author contribution, time difference between publications and dynamic combinations of indicators used.

The paper is structured as follows. In the next section, we review the literature on the current disambiguation methods being employed, with an emphasis on methods that mix both computer science, and sociological and linguistic approaches. We pay particular attention to the prevalence of data discarding and the effect it has on results, leading to the data and method section. The data and method section presents an overview of the strategy of the method, expectations of current methods (including our own), the data used, and data preparation. We explain in detail which meta-data objects are used and how these objects are employed in similarity calculations and logistic regression. We also explain how our method utilises two new meta-data objects—time between publications and author contributions—to achieve stronger disambiguation results. In the results section we apply our method to a test dataset to show its accuracy in terms of precision and recall. In the final section, we discuss the results and look forward to the next generation of methods for disambiguation.

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, which seems the more fruitful approach. 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. Kang et al. (2009) also use co-author web-based correlations and co-author-of-co-author (co-author expansion) techniques in their study. In disambiguating researcher names in patents, Raffo and Lhuillery (2009) investigate the different search heuristics and devise sequential filters to increase the effectiveness of their disambiguation algorithm. Song et al. (2007) have developed a two-stage approach to assist with the problem of disambiguating

persons with the same names on web pages and scientific publications; first using two topic-based models—Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA)—linking authors to words and topics, and then using a clustering method—hierarchical agglomerative clustering—to disambiguate names. The testing was conducted using web data and CiteSeer on a dataset of 750,000 papers.

The problem of ambiguity is also addressed in studies dealing with heterogeneous datasets, for example, linking patents to publications and authors to inventors (Cassiman et al. 2007; Meyer 2001; Raffo and Lhuillery 2009; Trajtenberg et al. 2006).

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 potentially biased in that records that did not exhibit any similarities in the cited references were treated as isolates and thus excluded from the final effectiveness results. Onodera et al.'s study is the most similar to ours in that they use similarity probabilistic techniques. They differ not only in their objective of disambiguation (as they aim to retrieve specific authors' documents, not to discriminate between different authors within a dataset) but also in the use of available meta-data fields. The fields they used include co-authorships, affiliation addresses, citation relationships, title words, interval in years between publications, author country, citation and co-citation relationships.

Data selection and discarding of records

The discarding of data seems to be a prevalent feature of the extant literature. In many cases, where only one information object is used in the analysis, it is logical that if that field is empty or null, the records cannot be used in the analysis. This results in discarded data and a loss of recall, and the lost data is not always addressed. Some examples of data discarding are discussed here.

Malin (2005) uses the peripheral social network of actors to disambiguate specific entities. Presumably most movies have at least one actor so this method may be sound for the particular data source. However, to translate this to the scientific sphere where co-authorships are used to disambiguate researchers, publications often have a single author and by discarding all publications with only one author is a drastic move. A recent study by Kang et al. (2009) use co-authorships in publications to disambiguate but make no mention of the issue of single-author papers in their methodology. Huang et al. (2006) rely on author meta-data as input for similarity calculations. However, the meta-data examples given are emails and URLS, and addresses and affiliations. The problems with the selection of these data are, for example, authors email addresses generally only include the corresponding author's email address, and physical addresses are typically not author specific (where, for example, 5 authors are present but only 2 addresses are given).

Data and 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 discriminate the pairings of nodes in terms of unique authorship.

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. However, to correctly assign a whole corpus of works to one person, using all misspelled variants of a person's name would dramatically increase the effort required to minimally increase the recall of that particular authors work. Furthermore, databases such as the Web of Science mobilize authors to correct meta-data, for example for misspelled authors' names. As authors have a vested interest in correctly spelled names, one may expect this type of mistakes to be increasingly solved.[1] As a result of this, we have chosen to discard any variants in the spelling of the last name, and will rely on one spelling of the name.

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 4,979 articles, letters, notes and reviews featuring 5,616 authors. The records were retrieved from Thomson Reuter's Web of Science (WoS) 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 assigned during the parsing process. There are 3,872 different last names and of these there are 2,014 last names which have more than one publication. There are 4,403 author last name and first initial variants, with 208 instances in which more than one author has the same last name and first initial and 366 authors who share their last name only with one or more other authors. We have focused our efforts on the instances in which there are more than one author with the same last name.

Data preparation

Each author/publication combination was assigned a unique identifier (U ID). This is to ensure that each and every author instance is regarded as unique at the beginning of the process. The contingent of meta-data present in each publication, and associated U ID, were marked. We have selected the following base meta-data from the available meta-data of WoS and provide an explanation for the choice and for the treatment of potential problems of the meta-data:

1. publication title words: title word choice by authors is generally considered to be related to content. Assuming the author is relatively consistent in his field, content (and thus title word choice) will remain relatively stable (Han et al. 2003) and the relative level of co-occurrence of title words between publications gives a strong indication of

[1] However, corrections are not always possible, partly because of the database structure. This holds for entries older than 1995 (email March 3, 2011, from Thomson Reuters).

whether Author *A1* is the same as Author *A2*. However, this may not be constant across fields or in fields with stylised titles. The changing lexicon and meaning of words may also play a part. Also, title words may have been chosen specifically to address a particular audience—the so-called audience effect (L. Leydesdorff 1989; Whittaker et al. 1989).

2. Publication abstract words: as with publication titles, word choice is related to content along with perceived application benefits and a general overview of the methodology and results. The additional data of application benefits and methodology gives a more detailed picture of the cognitive background of the work, which in turn gives more depth of information to the similarity comparison algorithm. With both title and abstract words, we removed stop words and stemmed words using SAINT's Word Splitter function (Somers et al. forthcoming).
3. Citations: Working within the same field, a researcher may base much of their work on specific previous studies in the field, adding to the unique 'characterisation' of their work. Citation behaviour is also punctuated by levels of self-citations, group citations and opportunistic citation (Aksnes 2003; Pasterkamp et al. 2007; Nicolaisen 2007) which only add to the characterisation of the citation list. It is this behaviour that allows citations to be regarded as an indicator of similarity. However, citations not only suffer from ambiguity themselves, but citation behaviour may be different between fields and therefore differently contribute to identification through similarity. We have chosen to use citations 'as is' and have not manually checked ambiguous citations.
4. Keywords: A publication generally contains both author generated keywords and journal indexer generated keywords that can be used to create a similarity measure between two publications (Matsuo and Ishizuka 2004). Author generated keywords may be more accurate reflections of the content rather than the indexers' keywords due to the "indexer effect" (Healey et al. 1986). Keywords (or more accurately, 'keyphrases') are normalized by removing spaces between words and by grouping highly similar keyphrases based on Damerau-Levenshtein edit distances (Damerau 1964; Levenshtein 1966).
5. Author listings: Researchers tend to co-author within their own field, generating co-authorship lists that do not diverge enormously from their home field. Co-authorship occurrences are not necessarily field dependent and when researchers do co-author, they tend to do so in the same topic areas repeatedly (Wagner-Döbler 2001). The higher the shared co-author count across different publications, the higher the likelihood that authors with the same name are indeed the same individuals. Co-author names are used in a last name, first initial format as not all records maintain a listing of all the authors' full names.
6. Author addresses: Addresses are commonly used in disambiguation studies as they may definitively link an author to an address and if two authors of the same name share an address the likelihood that they are the same person is high. However, the use of addresses is complicated by authors maintaining more than one address (guest lectureships etc.), by inconsistent spelling of addresses, incomplete addresses, no address given, or when multiple authors and multiple addresses exist on publication data (Tang and Walsh 2010).[2] The addresses are normalized for object order (for

[2] Various data repositories, namely WoS, are working to improve on the issue of multiple assigned author addresses, and newer publications in the database have direct indications of which author(s) links to which address(es).

example—house number followed by street name versus street name followed by house number) by using Damerau-Levenshtein distances. In the case of multiple addresses and multiple authors with no defined indication of author-address links, a probabilistic approach is used where each author on a publication has an equal probability of linking to any of the addresses presented.

7. Journal name: Research fields may be delineated by the set of core journals in which most publications are published. Assuming a level of consistency in researchers' chosen fields, the primary choices of which journal to publish in remain relatively constant (van den Besselaar and Leydesdorff 1996). However, changing journals in a field and any inter-, multi- or trans-disciplinary research output may not be targeted to a constant list of journals, resulting in a lower degree of similarity when comparing author publications (Loet Leydesdorff et al. 1994).

To complement the given meta-data fields, additional data are used in our similarity calculations. These are:

1. Difference in years between publications: 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, and with that, a change in popular co-authors, choice of title words and/or keywords and so on. By accounting for this time difference, the expected individual contributions of the base meta-data to discerning the probability of two publications being by the same author, may change as well.
2. Average author contribution: With indicators such as publication title, publication abstract, citations, and choice of journal—the selection of words, citations and journal is performed in various but typically unequal measures by the contributors in the publication (Yank and Rennie 1999; Bates et al. 2004). Therefore, when using the indicators it is necessary to take this inequality in contribution into account. For example, if a researcher is listed as 3rd or 4th author the probability that he has contributed heavily to word choice in the title or citations is lower compared to being 1st or 2nd author on the publication. Author contributions are calculated using the sum of the fractional author counts of the author positions of the two records using Moed's formula (2000). The contributions of the second and last authors are equal to 2/3 of the contribution of the first author. Any other authors contribute 1/3 of the first author. This is normalized so that the sum of all the fractions is equal to one.[3] For example, in a publication of 6 people, where a is the contribution of the first author (Moed 2000):

$$a + 2/3a + 1/3a + 1/3a + 1/3a + 2/3a = 1; \text{ and } a = 3/10$$

The author of a single-authored publication has maximum control over input, and from the formula of Moed—$a = 1$ and thus the maximum value for contribution is 1. The average author contribution measures the deviation from maximum input, i.e. how 'far' away an author is from the maximum. For each author pair being compared, the average distance from maximum of each author is the average author contribution (AAC), and this is on a scale of 0–1, where 1 signifies maximum input of the two authors being compared, that is to say that both authors are the only author in their respective publications.

[3] 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).

Null combination code (NC)

When each record is compared, the minimum shared available meta-data of each pair is referred to as the NC code. This "null data-field code" (NC) is a string of ascending order numbers where each digit signifies the presence of a valid field. For example, if only the title, labelled as "1", abstract –"2" and author assigned keywords –"4" are present the NC code will be 124.

Year difference (YD) categories

The YD is categorised as follows: (1) $\geq$2 years difference; (2) >2 and $\leq$5 years difference; (3) >5 and $\leq$10 years difference; (4) >10 years difference.

Similarity calculations

The similarity calculations are based on the Tanimoto coefficient—τ—and follow 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 meta-data fields for which the Tanimoto coefficient is calculated are[4]: (1) title words; (2) abstract words; (3) last names and first initials of co-authors; (4) cited references in whole-string form; (5) normalized author keywords; (6) normalized indexer keywords; (7) normalized research addresses; (8) journal names.

Logistic regression

Logistic regression requires the presence of two pre-determined groups. We start with identifying some of the authors' correct publications and some publications with the same author name that definitively belong to another group. With this, we created an input dataset in which the pre-determined groups are defined as group 2 (where the author/publication records being compared are definitively the same individuals) and group 0 (where the author/publication records being compared are definitively NOT the same individuals) as shown in Table 1.[5]

The independent variable (the meta-data fields) cells contain the raw similarity values of those independent variables. If the independent variable is not present to compare between U IDs, it is marked as being NULL. The NC code reflects which of the independent variables are present for each U ID pairing. The data was split into calibration and testing sets in an approximately 25:75 ratio to test the validity of the model. A regression was run with the NC codes as filters.

The full regression formula is as shown in Eq. 1. For each NC combination (where "Sim" refers to the degree of similarity between two author/publication pairs in a specific type of meta-data):

[4] The meta-data fields are compared across records that share the same last name only.

[5] To make the algorithm useful for completely unchecked sets and thus avoid excessive manual checking of records we are currently working on a sampling method which will be presented in a follow-up publication.

Table 1 Sample input data table for regression analysis

	Pairs		Independent variables			
Group	A	B	1	2	3	NC code
0	1	5	a	b	c	123
0	1	6	Null	b	c	23
0	–	–	–	–	–	–
2	2	3	Null	b	c	23
2	1	3	Null	b	Null	2
2	–	–	–	–	–	–

$$\ln(Y/1-Y) = \beta 0 + \beta_1(\text{SimCoauth}) + \beta_2(\text{SimAbstract}) + \beta_3(\text{SimTitle}) + \beta_4(\text{SimCitedRef}) + \beta_5(\text{SimAuthorKeywords}) + \beta_6(\text{SimIndexerKeywords}) + \beta_7(\text{SimRes.Address}) + \beta_8(\text{SimJournal}) + \beta_8(\text{AAC}) + \beta_8(\text{YDCategory}) \quad (1)$$

The β coefficients found in the regression are used to estimate the pairing probabilities of the unknown data set. The default decision rule threshold of 0.5 is used to determine calculated group membership. The flowchart in Fig. 1 summarises the order of operations in which the calculations are performed.

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 and identifies the communities of papers belonging to unique authors.

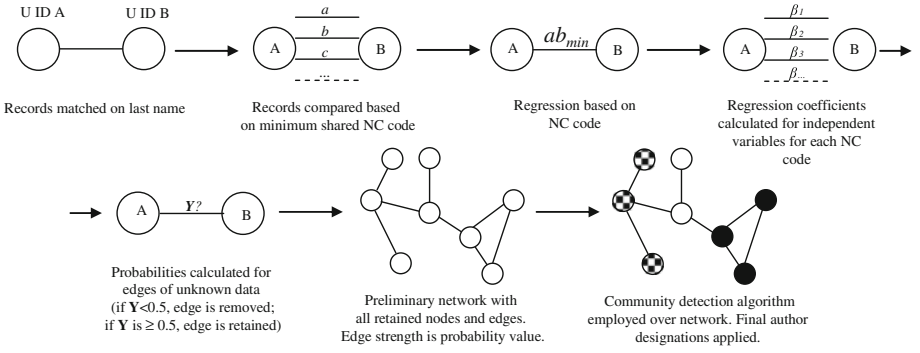

Fig. 1 Summary of order of operations of data processing, regression calculations and final author disambiguation

Results

To demonstrate the effectiveness of our algorithm we have chosen to present results based on matching last names only, and on matching last name and first initial. To demonstrate the importance of average author contributions and the time different between publications, we present the β values of the logistic regression for these variables from each NC code. The authors we have chosen as examples are all the authors from our dataset who have more than one publication and their last name is shared by one or more other authors. This list consists of 366 different authors with publication counts ranging from 2 to 420.

As a precursor to our results, we look critically at the potential problems of other methods, by using our own data of 366 different authors to demonstrate some of the fallibilities of alternative approaches. We have chosen to use the cited references as many previous studies use only cited references in their disambiguation efforts. We tested the number of cited reference matches between records and compared this to the mean number of cited references/publication. We have plotted the percentage of records which have one or no shared cited references and this is presented in Fig. 2.

In Fig. 2, as per the trend line, the mean number of cited references/publication does not deviate depending on how many publications an author has written. Contrary to this, the number of publications that share one or no cited references does vary with the number of total publications an author has written: the more publications an author has written, the fewer the publications that share one or no cited references. When disambiguating authors with relatively few publications, there is a much higher chance that the recall of their publications will be affected because there are fewer shared cited references. Due to the scarcity of shared cited references, the precision may also be negatively affected. This may be due to the exploratory nature of 'young' scientists' work. This is an important statistic to take into account when considering cited references as the primary source of meta-data for

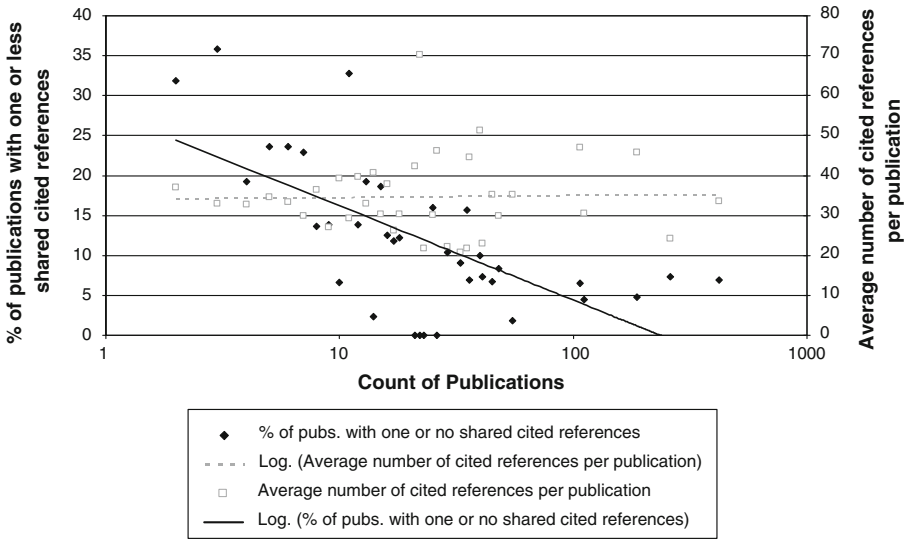

Fig. 2 Average percentage of publications with one or no shared cited references compared to average count of cited references per publication

disambiguation. It is difficult to build up a characterising aspect of an author if there are few or no similar characteristics between their publications.

For our primary results, we have used the harmonic mean version of the F-measure, with equal emphasis placed on precision and recall (Do et al. 2003). The F-measure (Eq. 2c) is composed of the precision and recall values as in Eqs. 2a and 2b:

$$\text{Precision (P)} = (\text{TruePositive})/(\text{TruePositive} + \text{FalsePositive}) \tag{2a}$$

$$\text{Recall (R)} = (\text{TruePositive})/(\text{FalseNegative} + \text{TruePositive}) \tag{2b}$$

$$\text{F} - \text{measure} = 2 \times (\text{PR}/\text{P} + \text{R}) \tag{2c}$$

We have calculated the average precision, recall and F-measure values on authors with varying counts of publications, based on using the last name, and last name and first initial.

Contributions of AAC and YD

Table 2 shows the β coefficients of AAC and YD to the logistic regression calculations. (A complete analysis of the logistic regression calculations including the β coefficients for the other independent variables will be presented in a follow up paper). For every NC code the AAC β is always higher than the YD β. The maximum possible value of the AAC is 1, signifying that the closer the two authors are to having maximum input on the publication, the higher the chances that the edge between the two publications in the network will be regarded as being a correct edge, i.e. the two publications are by the same person.

The further away the two authors are from maximum input, the lower the chances that the edge will be placed between the two publications. Of the variables available, the indexer keywords are not affected by the authors in any way. The research addresses are also not affected by the authors themselves as they are indicators of location rather than content. The variability of the AAC β when one takes into account what input authors have exactly on a publication is something to be investigated in the future. For YD, the β coefficients do not vary much over the different NC codes, which was unexpected as we theorised that the effect of time on similarity between publications would affect results more significantly. The results indicate that the effect of time difference between

Table 2 Contributions to group membership in logistic regression calculations by AAC and YD

C code	YD β	AAC β	NC code	YD β	AAC β	NC code	YD β	AAC β
357	0.14	2.484	23,579	0.128	1.277	234,579	0.146	1.212
1,357	0.2	2.551	34,578	0.102	1.828	345,789	0.134	0.856
2,357	0.073	1.784	34,579	0.128	0.97	1,234,578	0.142	1.722
3,457	0.184	2.411	123,457	0.137	1.873	1,234,579	0.182	1.236
12,357	0.125	1.911	123,579	0.174	1.292	1,345,789	0.17	0.881
13,457	0.22	2.481	134,578	0.14	1.887	2,345,789	0.153	1.096
13,579	0.154	1.128	134,579	0.166	1.055	12,345,789	0.186	1.065
23,457	0.1	1.766	234,578	0.107	1.644			

The NC code signifies the available meta-data objects. The presence of each number signifies the presence of a specific meta-data object. The numerical codes for each object are: *1* co–authorship, *2* abstract, *3* title, *4* cited references, *5* journal, *7* research address, *8* author keywords, *9* indexer keywords (*6* journal Category is not shown)

publications decreases when the abstract is included in the analysis. Abstracts seem to have a larger similarity over time through a recurring use of some words.

Results based on last name only

Figure 3 shows the average precision, recall and F-measure of authors of varying publication counts. The distribution of the number of authors with specific counts of publications is expected, i.e. there are many authors who have published little, and few authors that have published prodigiously. The average recall values/publication total are all above 0.85. The precision values are mostly higher than 0.9 with a few exceptions. Almost all the F-measure values are above 0.8 with most above 0.9, with one exception at 0.5.

The exceptions to these scores are primarily due to two different authors with the same last name but different first initial, being incorrectly designated as a single author, in which one of the authors has an exceptionally high count of publications ($\sim$200) and the other a relatively low count of publications ($\sim$20). A similar situation which affected the F-measure occurred when one author has been deemed to belong to two communities, i.e. the algorithm has classed the one author as two separate authors. In the case of this happening, we have used the average of the "two" authors to give a single result.

Overall, the results based on last name only are very high, which given the number of authors, and the count of authors with the same last name, is very good.

Results based on last name and first initial

Figure 4 shows the average precision, recall and F-measures of the same authors which were presented in Fig. 3, but now on a last name and first initial level. The distribution of number of authors with specific publication counts remains the same. Compared to Fig. 3, the results using last name and first initial are, as expected, better is almost all aspects. The problem of two authors with the same last name but different first initials being incorrectly classed as the same author has been removed. The low result at 0.5 in both figures is from one author who has been incorrectly designated as two separate authors. On overview of

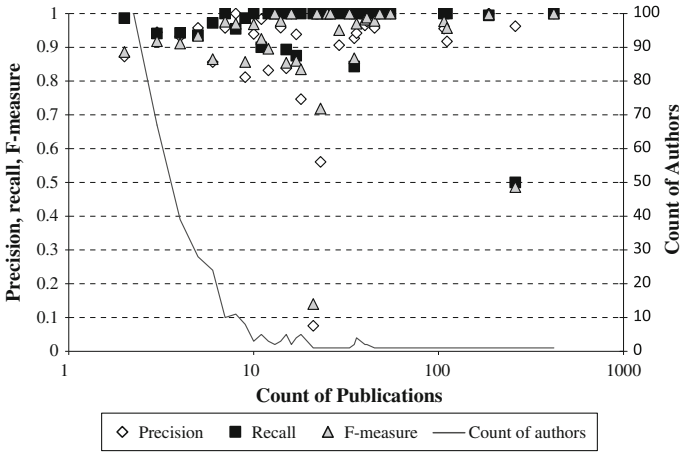

Fig. 3 Average recall, precision and F-measure values including author publication count distribution using last name only

Fig. 4, the results for each publication count category of authors, has increased as compared to Fig. 3. There are many more perfect scores (P = 1, R = 1, F-measure = 1) over many more publication count categories.

To summarise the results as seen in Figs. 3 and 4, the algorithm has worked remarkably well across the range of authors with different publication counts. This is important as it shows the algorithm is extremely suited to discerning authors who have few publications who may not publish repeatedly in the same field. An example of this may be young scientists' PhD-related publications, as compared to their post-doc or further publications.

For authors who have many publications, the algorithm works well to assign these publications to the correct author, who may have changed research topics multiple times over the years.

Discussion and conclusions

Author disambiguation will be an ongoing problem for some time, even with database providers working to solve the problem. The move towards placing the onus of identification on authors may be a step forward. But the records of authors who are no longer active in publishing may remain ambiguous for the foreseeable future. It is for this reason that algorithms such as ours will remain important for researchers who make use of bibliometric data.

Our choice to compare records based on the last name and first initial, and on last name only was a result of our need to test the discerning power and robustness of our algorithm. By creating a very large number of possible matches (our final master table of potential match records had over 1.5 million rows) we intended to stress our computing power, and the ability of the algorithm to manage this large number of records.

Our method differs to previous methods in three ways. Firstly, we do not pre-select records based on specific meta-data. Rather we utilise every meta-data object available. The retention and utilisation of all possible meta-data has proven to be helpful as records that did not display any similarity on, for example, cited references may still have shown

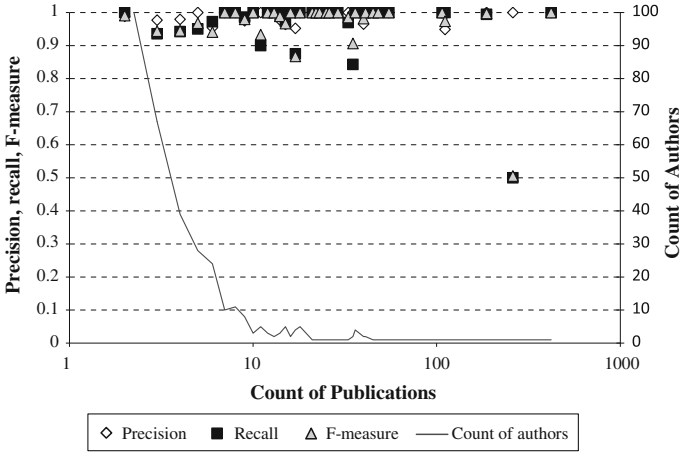

Fig. 4 Average recall, precision and F-measure values including author publication count distribution using last name and first initial

high topic similarities as seen in the abstract or title words. A key fact of the data we examined was that there were a substantial proportion of records missing meta-data. The variability of what meta-data was available to compare, spurred us to think of a dynamic approach in which only the minimum shared meta-data was utilised. This meant that a record could be compared to others on completely different meta-data for each comparison. The use of logistic regression was required for this as we wanted to know the contributions of each data object to discerning group membership, and we realised that for each combination of available data objects, there would be different levels of contribution by each object.

Secondly, the additional meta-data that we have chosen to include, those of time difference between publications and average author contribution, have been important. The goal of our disambiguation method (and that of many other similar methods) is to create a continuous chain of publications—a coherent subnetwork within the larger network. By accounting for the age difference between publications we increase the chances of young publications linking to older ones, not just similarly-aged publications linking to each other. Changing subjects and influences of a researcher over time create a longitudinally stretched network of publications which, when thresholds are applied, are susceptible to being broken. By linking the older and younger publications, we increase the chances that the subnetwork remains intact. The average author contribution meta-data was very important in that it also gave the algorithm room for flexible similarity parameters. Tang and Walsh (2010) mention the fact that other authors in a publication have an influence on what meta-data is included in the final version of the publication, thus affecting the "knowledge homogeneity" of the author under inspection. We have successfully shown that recognising and, more importantly, using this difference in author contribution actually increases the coherence of the subnetwork of publications of a specific author. Together, these two additions to the range of employed meta-data increase the deductive power of the algorithm.

The retention of all possible meta-data has also proven to be helpful as records that did not display any similarity on one variable, for example cited references, may still have shown high topic similarities in other variables, such as the abstract or title words. 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.

Previous studies commonly use thresholds to increase accuracy rates, which are useful in a proof-of-concept, but in real situations there is no way to know which threshold is the best to use. Our method does not use any thresholds, apart from the default 0.5 threshold for logistic regression which, when translated to real-world operations, is far easier to manage and replicate for further studies.

To move our algorithm from proof-of-concept to working process, we need to address the issue of pre-checking records. There is a substantial amount of manual work involved in all methods (including ours). (We are currently working on a method that reduces the manual work involved substantially and this will be presented in a follow up paper.). At present, excluding the previous authenticity checks performed by the originators of the dataset, the method—from parsing publications to final author designation—takes approximately 8 h, of which the most time is spent importing the logistic regression results from SPSS into Access. The use of a plug-in for R (an alternative statistical analysis program) is being investigated which would reduce the time spent immensely.

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 precision and recall. With the benefit of further research, we will investigate the minimum number of publications necessary to consistently and accurately disambiguate authors.

To summarise, our method retains all data and discards no information, accounts for activity of authors in different fields or specialties (year difference) and in different capacities (AAC); uses no arbitrary thresholds; is scalable; and provides highly accurate disambiguation results.

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 or inventor on the order of hundreds of thousands of records.

References

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics, 56*(2), 235–246.

Bates, T., Anic, A., Marusic, M., & Marusic, A. (2004). Authorship criteria and disclosure of contributions: comparison of 3 general medical journals with different author contribution forms. *JAMA, 292*(1), 86.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Cassiman, B., Glenisson, P., & Van Looy, B. (2007). Measuring industry-science links through inventor-author relations: a profiling methodology. *Scientometrics, 70*(2), 379–391.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171–176.

Do, H. H., Melnik, S., & Rahm, E. (2003). Comparison of schema matching evaluations. *Web, Web-Services, and Database Systems, 2593*, 16.

Gurney, T., Horlings, E., & Van Den Besselaar, P. (2011). Author disambiguation using multi-aspect similarity indicators. In: E. Noyons, P. Ngulube, J. Leta (Eds), Proceedings of ISSI 2011—The 13th International Conference on Scientometrics and Informetrics, Durban, 4–7 July 2011, pp 261–266.

Han, H., Zha, H., & Giles, C.L. A model-based k-means algorithm for name disambiguation. In, 2003: Citeseer.

Healey, P., Rothman, H., & Hoch, P. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233–251.

Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. *Lecture Notes in Computer Science, 4213*, 536.

Kang, I. S., Na, S. H., Lee, S., Jung, H., Kim, P., Sung, W. K., et al. (2009). On co-authorship for author disambiguation. *Information Processing & Management, 45*(1), 84–97.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization 1. *Research Policy, 18*(4), 209–223.

Leydesdorff, L., Cozzens, S., & Van den Besselaar, P. (1994). Tracking areas of strategic importance using scientometric journal mappings. *Research Policy, 23*(2), 217–229.

Malin, B. Unsupervised name disambiguation via social network similarity. In, 2005 (pp. 93–102): Citeseer.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools, 13*(1), 157–170.

Meyer, M.S. (2001). *Patent citation analysis in a novel field of technology: an exploration of nano-science and nano-technology* (Vol. 51, pp. 163–183). Berlin: Springer.

Moed, H. F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics, 47*(2), 323–346.

Moed, H. F., Glänzel, W., & Schmoch, U. (Eds.). (2004). *Handbook of quantitative science and technology research, The use of publication and patent statistics in studies of S&T systems*. Dordrecht: Kluwer Academic Publishers.

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology, 41*(1), 609–641.

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani, Y., 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.

Pasterkamp, G., Rotmans, J. I., de Kleijn, D. V. P., & Borst, C. (2007). Citation frequency: a biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics, 70*(1), 153–165.

Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117–136.

Raffo, J., & Lhuillery, S. (2009). How to play the "names game": patent retrieval comparing different heuristics. *Research Policy, 38*(10), 1617–1627.

Somers, A., Gurney, T., Horlings, E., & Van Den Besselaar, P. (2009). *Science assessment integrated network toolkit (SAINT): a scientometric toolbox for analyzing knowledge dynamics*. The Hague: Rathenau Institute.

Song, Y., Huang, J., Councill, I.G., Li, J., & Giles, C.L. 2007. Efficient topic-based unsupervised name disambiguation. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007) (pp. 351). New York: ACM.

Tan, Y.F., Kan, M.Y., & Lee, D. (2006). Search engine driven author disambiguation. In 6th ACM/IEEE-CS joint conference on Digital libraries: Chapel Hill: ACM.

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics, 84*(3), 763–784.

Trajtenberg, M., Shiff, G., & Melamed, R. (2006). *The Names Game: Harnessing Inventors' Patent Data for Economic Research*. Cambridge: NBER working paper.

van den Besselaar, P., & Leydesdorff, L. (1996). Mapping change in scientific specialities; a scientometric case study of the development of artificial intelligence. *Journal of the American Society of Information Science*, *47*(5).

Wagner-Döbler, R. (2001). Continuity and discontinuity of collaboration behaviour since 1800—from a bibliometric point of view. *Scientometrics, 52*(3), 503–517.

Whittaker, J., Courtial, J. P., & Law, J. (1989). Creativity and conformity in science: titles, keywords and co-word analysis. *Social Studies of Science, 19*(3), 473–496.

Yang, K.H., Peng, H.T., Jiang, J.Y., Lee, H.M., & Ho, J.M. (2008). Author Name Disambiguation for Citations Using Topic and Web Correlation. *Research and Advanced Technology for Digital Libraries*, 185–196.

Yank, V., & Rennie, D. (1999). Disclosure of researcher contributions: a study of original research articles in The Lancet. *Annals of Internal Medicine, 130*(8), 661.

Zhu, J., Zhou, X., & Fung, G. (2009). A term-based driven clustering approach for name disambiguation. *Advances in Data and Web Management*, 320–331.