Scientometrics
DOI 10.1007/s11192-010-0196-6

Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps

Li Tang · John P. Walsh

Received: 25 September 2009

Abstract Authorship identity has long been an Achilles' heel in bibliometric analyses at the individual level. This problem appears in studies of scientists' productivity, inventor mobility and scientific collaboration. Using the concepts of cognitive maps from psychology and approximate structural equivalence from network analysis, we develop a novel algorithm for name disambiguation based on knowledge homogeneity scores. We test it on two cases, and the results show that this approach outperforms other common authorship identification methods with the ASE method providing a relatively simple algorithm that yields higher levels of accuracy with reasonable time demands.

Keywords Name disambiguation · Common names · Cognitive map · Approximate structural equivalence · Knowledge homogeneity score · Hierarchical clustering

Introduction

Authorship uncertainty is a ubiquitous challenge for many fields ranging from art museums to credit bureaus to crime investigation (Borgman and Siegfried 1999; Chaski 2005; Pasula et al. 2004; McCallum and Wellner 2003). In the arena of bibliometrics, it also has been a classic problem (Garfield 1969), but for a long time the literature has struggled with hand methods or, worse yet, ignored this issue. For example among 515 articles in *Scientometrics* published between 2006 and 2009,[1] only two articles explore this issue (Soler 2007;

[1] The search was conducted on July 26, 2009.

Both authors contributed equally to this work and are listed alphabetically.

L. Tang · J. P. Walsh (✉)
School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: jwalsh6@mail.gatech.edu

L. Tang
e-mail: litang@gatech.edu

Published online: 08 April 2010

Wooding et al. 2005). Some studies simply avoid micro-level analysis with bibliographical data, some indicate a method without elaboration on how the issues are dealt with, while others show the results of analyses, but keep authorship identification as a black box.

But authorship identification matters a great deal and has to be taken seriously for any analysis at the individual level for two main reasons. First, the fast expansion of the number of researchers and the rise of large-scale digital libraries are making existing methods that depend on hand-checking cases increasingly less tenable. Second, we have seen an increasing internationalization of science and especially the rising prominence of China in the global science system. From 1995 to 2005, China's share of scientific publications increased from fourteenth place to fifth overall, second place in engineering and chemistry and third place in physics and mathematics (NSF, 2008). Recent data suggest that China may have moved into the number two position, behind only the US (Zhou and Leydesdorff 2008; Kostoff 2008). The rise of China has made disambiguation even more difficult, because of the large number of Chinese scholars sharing a few family names such as Zhang, Wang, Li and Chen (Strotmann et al. 2009).

To address the name disambiguation problem, several approaches, particularly from the information science field, have been proposed. The common names problem, the most difficult challenge in name disambiguation, however, remains unsolved. In this paper, we target on common names of researchers, and propose an Approximate Structural Equivalence (ASE) algorithm to trace authors' bibliometric fingerprints based on their knowledge homogeneity scores. Experiments on two sets of predefined articles: one a relatively common American name using data from the Social Science Citation Index and another a common Chinese name using data from nanotechnology papers drawn from the Science Citation Index, demonstrate the effectiveness of this proposed approach. In comparison with other name disambiguation methods, the ASE method is better suited to large datasets, is easily understandable, is less time consuming, and has a higher accuracy rate. Additionally, since it uses clustering algorithms rather than pair-wise matching, and it does not depend on affiliation information, the ASE approach can easily track mobile researchers or inventors and lends itself to automation. Its insensitiveness to name variations allows for tracking authors who change their names, for example, because of marriage. It is also useful for distinguishing authors within a given field, while other similar algorithms may have more difficulty distinguishing them based on key words, technology classes or collaboration patterns (Raffo and Lhuillery 2009).

The rest of this paper is organized as follows. We start with a review of the problems of authorship identification in bibliometric analysis and the extant methods. Then we introduce our proposed ASE method. We further test this approach on two examples and benchmark it against other methods. We conclude with a discussion of contributions and limitations of this approach.

Literature review

The name ambiguity problem

The key issue of name ambiguity is to make certain whether two archival records with the "same" or "similar" names refer to the same researcher.[2] This simple task remains as an

[2] Using the example of patent documents, Trajtenberg illustrates this problem with two questions: (1) Is "Manuel Trajtenberg" in one patent the same inventor as "Manuel Trajtenberg" in another record? And (2)

Table 1 Examples of name ambiguity problems

Factors	Example	Consequences
Variations of personal spellings	Walsh, J Walsh, J P John P. Walsh J P Walsh	Type I error of undermatch
Typographical and phonetic errors	Wlash, P Walhs, P	Type I error of undermatch
Translation and transliteration	Li Yue is the translation of different Chinese names such as “李越”, “李月”, “李跃”, “黎悦”, “厉乐”, to name just a few, as well as any combination of these first and second Chinese characters Li Yan could be “李妍”, “李燕”, “李彦”, “李岩”, “李延”, “李炎”, etc.	Type II error of overmatch
Name changes over time	Anne Walsh changed her name after marriage	Type I error of undermatch
Common names	Smith, Walsh, Li, Kim, etc.	Type II error of overmatch

unsolved fundamental problem in bibliometric analysis for numerous reasons. To begin with, a single researcher may be associated with different names due to: (a) variations of personal spellings, (b) typographical and phonetic errors, (c) translation and transliteration, and (d) name changes over time associated with marriage and other reasons. For a good summary please refer to Smalheiser and Torvik (2009) and MacRoberts and MacRoberts (1989). A second concern, which has been the focus of much recent research, is different individuals with the same names, particularly common names. This becomes increasingly problematic with the growth of the number of active scientists. In addition, the emergence of interdisciplinary research and research collaboration make research subject and affiliation rather weak to aid author differentiation. A third factor is related to the limitations of some commonly used publication databases, which mainly serve for research content retrieval and are relatively weak in authorship identification. As an example, in the ISI Web of Science, a very widely used standardized digital library, the full names of authors are not available for papers published before 2007. In addition, for multi-authored, multi-affiliation papers, once the publication records are downloaded, one cannot always be sure about the match between affiliation and authorship except for the correspondent author. These factors intertwine with each other and make author name disambiguation a big challenge for bibliometric analysis at the micro-level. Table 1 gives examples for each situation and lists the types of error incurred if a specific author is targeted.

Why it matters

Research on authorship identity has shifted from a period of viewing this as not a “major problem” (MacRoberts and MacRoberts 1989) to increasing interest in this topic (Houvardas and Stamatatos 2006; Smalheiser and Torvik 2009). Phelan (1999) argues that this

Footnote 2 continued
is “Manuel Trajtenberg” the same inventor as “Manuel Trachtenberg”, and the same as “Manuel D. Trajtenberg”? (Trajtenberg et al. 2006).

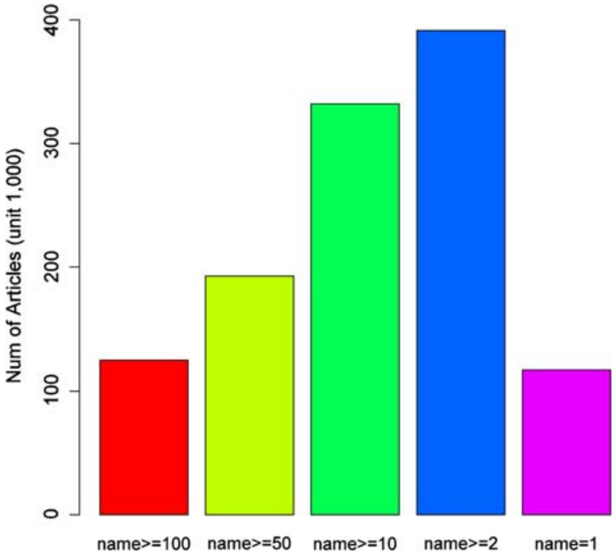

Fig. 1 Frequency of author names in nano-research, 2000–2007 (401,381 papers total)

name ambiguity problem is far more common and influential than generally imagined (Phelan 1999). Strotmann and his colleagues make a compelling argument on how collaboration networks look significantly different with and without authorship identification (Strotmann et al. 2009). Following Strotmann, we begin with a quick inspection into nanotechnology research as an illustrative case.[3] We find that the number of research papers containing identical names in this domain is substantial. Out of approximately 400,000 nanotech papers published in 2000–2007, over 390,000 records, i.e., 98%, are associated with names that appeared more than once (Fig. 1). Over 80% of articles contain author names that appeared at least ten times. Almost half the articles are related to very common names, or very prolific authors, each appearing more than 50 times in this period, with an increase in common names over this time period.

A closer examination also suggests that ignoring the name ambiguity problem would introduce significant biases. Table 2 lists the top twenty author names appearing in the nano paper database.[4] Excluding "Anon" (Anonymous), the most prolific authors are disproportionately associated with Chinese family names. This conveys two messages: (1) common names are not orthogonal to country and thus cannot be thought of as "random errors"; and (2) analyses on Chinese researchers, such as studies of productivity, research quality, collaboration networks and so on, need to be especially cognizant of this problem.

[3] The figures here are calculated based on a unique nanotechnology publication database developed by a Georgia Institute of Technology research group led by Dr. Philip Shapira. For a detailed description of this dataset, please refer to Porter et al. (2008).

[4] These author names are taken directly from the nano publication dataset derived from WoS without any cleaning. So although possibly part of "Kim, J" is overlapping with "Kim, J H", they are taken as they were in this table.

Table 2 Top twenty common names[a] in WoS nano publications: 2000–2007

Rank	Author name	Number_records	Family-name-based author origin
1	Anon	1847	Anonymous names
2	Wang, J	1265	Chinese family name
3	Zhang, Y	1232	Chinese family name
4	Wang, Y	1116	Chinese family name
5	Liu, Y	975	Chinese family name
6	Li, Y	932	Chinese family name
7	Zhang, J	920	Chinese family name
8	Chen, Y	849	Chinese family name
9	Li, J	843	Chinese family name
10	Wang, H	814	Chinese family name
11	Wang, L	800	Chinese family name
12	Zhang, L	747	Chinese family name
13	Wang, X	746	Chinese family name
14	Kim, J	736	Unspecified
15	Liu, J	711	Chinese family name
16	Kim, J H	639	Unspecified
17	Chen, J	633	Chinese family name
18	Lee, J H	626	Unspecified
19	Zhang, X	585	Chinese family name
20	Zhang, H	569	Chinese family name

[a] There are two Pinyin systems in Chinese names. The Mandarin Pinyin, or Hanyu Pinyin, is adopted in the mainland. Taiwan and Hong Kong Special Administrative Regions use Tongyong Pinyin system. The differences between the two Pinyin systems are less than 20%. In this article, we are using the Hanyu Pinyin system

Existing methods

For a very long time, hand checking has served as an acceptable (though time consuming) method for name disambiguation. However, given the magnitude of the problem discussed earlier, manual work is not only tedious, but as the database grows, may be prohibitively time consuming. Besides hand checking, other existing methods on authorship identification can be categorized into the following groups. The first group, a laissez faire method, assumes the errors are randomly distributed and simply ignores this problem. In this method, author names that exactly match each other are taken as the same author, otherwise not.[5] This method is rather quick, but rarely adopted nowadays.

The second approach takes us a step further. Acknowledging different spelling formats and typos in archival databases, a fuzzy matching method is adopted based on the similarity of reported author names, with or without affiliation verification or research subject verification. Some text mining software such as VantagePoint[6] has developed this function

[5] For example, in their report gauging the structure and competitiveness of China's nanotechnology, Kostoff and his colleagues identify "Zhang, Y." and "Li, Y." as the top two most prolific Chinese nano authors in 2003 (Kostoff et al. 2006, p. 148), ignoring the fact that Zhang and Li are also the top two most common family names in China.

[6] Text mining tool developed at Search Technology, Inc.

to facilitate matching records (Youtie et al. 2008; Frietsch et al. 2008). Studies adopting this approach often report tremendous efforts on data standardization without elaborating on how decisions were made on common names, researcher mobility, missing data on affiliation, as well as interdisciplinarity issues. For example, consider the "name + affiliation" algorithm, one common method for name disambiguation. This method would under-estimate the productivity and collaboration network scope of mobile researchers. Even more problematic, the applicability of this approach is limited by data availability. Without going through each original full text article or at least the original cover page, this method can be readily applied only if: (1) the targeted author is the single author, or (2) the targeted author is the reprint author, or (3) only one affiliation is reported in that article. Otherwise, we cannot uniquely assign each author to his/her institution(s) in multi-institution cases, an increasingly common occurrence (Jones et al. 2008). In addition, considerable numbers of typographical errors and organization name changes have been found in affiliation name and related geographical information. This is particularly true for non-English publications due to translation and transliteration. Take China for example. Peking University sometimes is also translated into Beijing University; Chinese Academy of Sciences is abbreviated as CAS, or Academia Sinica[7]; and Zhongshan Medical University changed its name to Zhongshan University after being merged in 2001, and is also commonly referred as Sun Yat-sen University, to name just a few of the common problems with institution names.

The third stream of disambiguation methods uses multistage matching. For example, Trajtenberg et al. (2006) proposed a two-stage model for inventor name matching. In the first stage, all "suspected" records with identical or sufficiently similar names were put together and coded via Soundex.[8] In the second stage, any link between the two records (address, technical field, assignee shared "partner", cite each other, etc.) are given scores if the link exists. Then overall scores are computed. If the score is above a threshold, a "match" decision is made (Trajtenberg et al. 2006). Similarly but with more sophisticated matching algorithms, Raffo and Lhuillery (2009) compare simple string matching in the first stage (N-gram, Token, with and without weighting), and then use a multiple filtering algorithm based on commonly available elements in patent documents (location, technical field, assignee name, cross-citation) to disambiguate names. They find significant gains in precision from the use of multiple filters. The problem of the above methods is it is extremely time-consuming to compare each pair of records and compute the overall score based on each matching criterion. In addition, accuracy depends on which algorithms and filters are used, and so it is not clear which specific criteria should be used on a particular type of data. Other advanced methods such as the "Author-ity" model developed by Smalheiser and Torvik for MEDLINE publication (2009), co-inventor networks approach to identify U.S. patent-holders (Lai et al. 2009), support vector machines for cyber authorship (Abbasi and Chun 2006; Huang et al. 2006), N-Gram feature on stylometry (Houvardas and Stamatatos 2006), authorship network (Wooding et al. 2005), linear discriminant function analysis (Chaski 2005), random forest model (Treeratpituk and Giles 2009), K-cluster (Han et al. 2005), Hierarchical Naive Bayes Mixture Model (Han et al. 2005) and supervised machine learning (Han et al. 2004) have also been proposed for different research settings. A majority of these types of studies come from information and

[7] Academia Sinica also exists in Taiwan.

[8] The Soundex algorithm transforms names into alphanumeric codes targeting on the variations of name spellings. This method was initially developed by the US Census in 1930. To apply this method, a full name and, preferably, residing state, must be known. For more details on this index method please refer to http://www.archives.gov/genealogy/census/soundex.html.

computation science fields. For a good summary please refer to Kang (2009).[9] Similar to Trajtenberg's approach, these methods usually require comprehensive data collection and coding, and often suffer from missing data problems. These methods also require access to unpublished algorithms and decision rules. This is one reason we could not replicate these methods in our two cases for benchmarking.

Method

Basic idea

Because of the limitations of existing methods, but still building on the lessons learned from these efforts, we propose an alternative method stimulated by the concept of cognitive map in psychology and structural equivalence in network analysis. We begin with the assumption that research papers are a reflection of the knowledge base of the author(s), with each author drawing from his or her own knowledge base that is generated through his or her particular training and experience. In particular, coming from different fields, subfields, and institutions is likely to expose one author to a different set of published and unpublished literature from another. Similarly, attending particular conferences and workshops is likely to give one access to a specific set of unpublished or recently published papers. When writing their own papers, authors draw on this unique collection of acquired research results. That is to say, within a certain period, the same author is drawing from the same knowledge set, while different authors (with the same name) draw on different knowledge sets. This process of acquiring, storing, and recalling knowledge and experience is similar to mental representation of allocentric space in individual's cognitive mapping (O'Keefe and Nadel 1978; Jacobs and Schenk 2003).

To visualize the linkage of different articles written by one author, we further borrow the concept of structural equivalence (SE) in social network analysis. Briefly speaking, in a single-relation network, two actors are structurally equivalent if they have identical ties to and from all the other actors (Lorrain and White 1971; Hanneman and Riddle 2005; Wasserman and Faust 1994). In reality, however, true structural equivalence is rare. Therefore, the definition of equivalence is relaxed to approximate structural equivalent (ASE), such that actors within a structurally equivalent cluster are more similar to each other than to those outside the cluster. Applying this notion to the name ambiguity issue, two articles are considered approximately structural equivalent if they are similar in position for referencing article(s) in an article-reference bipartite network (Pieters et al. 1999). If these structurally equivalent records contains author names with the same (or similar) family name and first initial, these similar author are taken to be the same authors.

Figure 2 illustrates this idea. Two research papers AR_1 and AR_2 have their own set of references, while some of these references are common (such as Ref_p). If those shared reference(s) are important enough to suggest a high knowledge similarity between AR_1 and AR_2 (i.e., share a certain number of common references or a rare reference, see below), these two *papers* are approximately structural equivalent (ASE). If the authors of ASE

[9] In addition to the above methods, large digital libraries have themselves also started to build up author identification systems, such as the Distinct Author Identification System (DAIS) in WoS (http://science.thomsonreuters.com/training/wos/), Citeseer (http://citeseer.ist.psu.edu/), and Research ID system (http://www.researcherid.com/), respectively. Unfortunately, our experiments disclose that their performances are rather poor, at least in the two common names we tested.

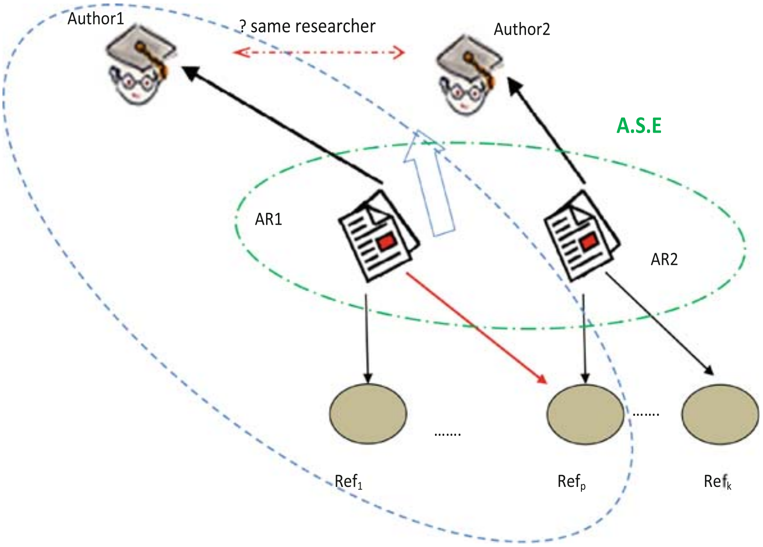

Fig. 2 Authorship identification based on shared references (the graph is adapted from Brian Milch's presentation "Relational Probability Models" at IPAM Summer School, 2007. The electronic version is available at: http://www.ipam.ucla.edu/publications/gss2007/gss2007_7224.pdf)

papers (Author 1 and Author 2 in Fig. 2) have the same family name and first initial, it is highly probable they are the same author.[10]

ASE algorithm

There are two subtasks associated with name disambiguation. One is to decide who wrote the specific targeted paper and the other is to find out which papers are written by this researcher (Smalheiser and Torvik 2009). The ASE algorithm address these two subtasks by partitioning authors into blocks based on their reference networks.

The unit of analysis of our ASE algorithm is the similarity of each pair of articles, which is measured by the knowledge homogeneity score (hereinafter KHS). The value of the KHS is determined by three factors: the summation of shared references, the forward citations of each reported reference, and the minimum number of references reported by the two articles. We assume a researcher's knowledge stock on a specific research problem can be reflected by the reference coverage of that paper at a given time. The more references two articles share, the more likely they are written by the same author. Given the heterogeneity of cited references, two weights $\mathbf{W_1}$ and $\mathbf{W_2}$ are used to moderate the predictability of shared references on article clustering. Intuitively, the more famous the cited reference is, the more likely that it is cited by different researchers.[11] In contrast,

[10] Please note approximately structural equivalent is different from co-citation, another similarity measure used in bibliometric analysis. The former focuses on records sharing references, while the latter refers to citations themselves. In the Fig. 2, AR1 and AR2 are ASE while ref$_1$ and ref$_p$ are co-citations.

[11] For example many sociologists cite The Sociology of Science (Merton 1973), and science historians often cite The Structure of Scientific Revolutions (Kuhn 1962), and network methods papers often cite Social Network Analysis (Wasserman and Faust 1994).

the chance that two different researchers cite a newly presented conference paper would be very low. So if two articles are associated with an author of the same family name and first initial, and both of them reference a rarely cited article, the chance that these two articles are written by the same author will be extremely high. This is analogous to the fingerprint biometric system where the fingerprint tail differentiates people.[12] In terms of number of references reported, the more references that an article reports, the higher probability it shares references with the others. Thus, if two articles report a small number of references but still share a certain number of references, the chance that they are contributed by the same author will be much higher than articles reporting many references.

Mathematically, the knowledge homogeneity score, i.e. KHS matrix can be denoted as:

$$\mathrm{KHS} = [\mathbf{A} * \mathbf{R}]_{n\times m} * \mathbf{W_1}_{m\times m} * \mathbf{W}'_{\mathbf{1}m\times m} * [\mathbf{A} * \mathbf{R}]'_{m\times n} \bullet \mathbf{Q}_{n\times n}$$

where **A** is one set of targeted n publications, **R** is the set of shared m references reported by **A**. $[\mathbf{A} * \mathbf{R}]_{n\times m}$denotes a two-mode, unidirectional co-occurrence matrix, in which the value of each cell is either 1 or 0 based on whether **A** cites **R** or not.

$$\mathbf{W_1}_{m\times m} = diag(\mathbf{W_1})\begin{pmatrix} w_{11} & & & \\ & w_{12} & & \\ & & \circ & \\ & & & w_{1m} \end{pmatrix},$$

where $\mathbf{W_1} = \{w_{11}, w_{12},\ldots, \mathrm{w}_{1m}\}$ is an ordered weighting vector with a dimension of $1 * m$, and the off-diagonal elements of $\mathbf{W_1}_{m\times m}$ are all set to 0. Its value is based on the number of forward citations of each reference, and $\mathbf{W}_{1j}$ is the weight of the jth reference, $1 \le j \le m$.

$\mathbf{Q}_{n\times n}$ is a weighting matrix transformed from $\mathbf{W_2}$ with $q_{ij} = q_{ji} = \max\,(w_{2i}, w_{2j})$, where $\mathbf{W_2} = \{w_{21}, w_{22},\ldots, \mathrm{w}_{2n}\}$ is an ordered weighing vector based the number of references, and w_{2j} is the weight of the jth article, $1 \le j \le n$, • denotes the entrywise product.

Once the KHS matrix is constructed, hierarchical clustering with *single* linkage is adopted to differentiate groups. If the knowledge homogeneity score between article i and article j, i.e., KHS[i, j] is above a KHS threshold, and KHS[i, k] is also above the KHS threshold, then article i, j, k will be grouped together as written by the same researcher.

Similar to other name matching methods, the shared-reference based clustering algorithm is vulnerable to two types of mismatching. Type I error of under-match could occur if a researcher has such broad research interests that one of his papers does not have any overlapping citations with the rest of his articles. On the other hand, a Type II error could happen if two researchers focus on the same topic and thus read the same literature. The weightings try to reduce both types of errors by setting up a sufficiently high threshold while weighting rare references more heavily. Our method also accounts for within-author shifts in subject area across papers by using a hierarchical agglomerative method with nearest neighbor criterion. This aims to impose transitivity, allowing for the fact that a researcher may not focus on one research topic within a given period. According to the principle of "friends of friends", if the knowledge homogeneity score determines AR_1 and

[12] In the arenas of criminology and anti-terrorism war, fingerprints tracing has been used to identify individuals (Chaski 2005) given that an individual has a unique finger *tail* that distinguishes her/him from the others. In the same vein, assuming an individual researcher has a fixed knowledge stock during a given period, tracing his/her bibliometric fingerprint in reported references should be useful for aiding authorship identification at a very large scale.

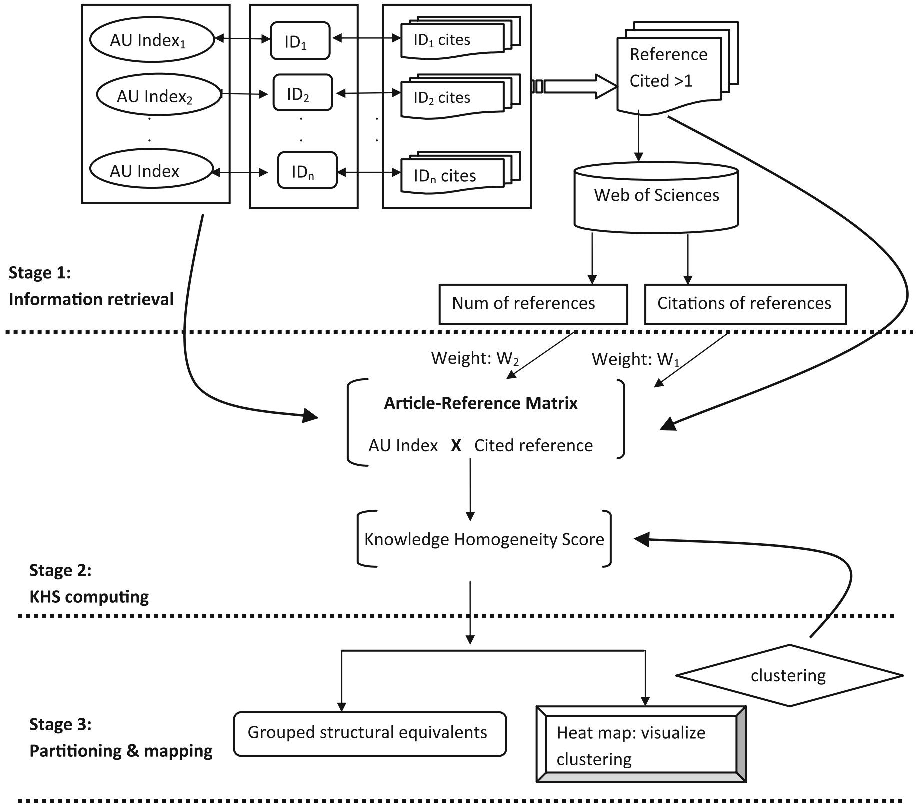

Fig. 3 Flow chart for author identification

AR_2 are written by the same author, while AR_2 and AR_3 are written by the same author, then AR_1 and AR_3 share an author even if these two research papers themselves have no shared citations. In other words, any author will be placed in one and only one cluster.[13]

The whole process is illustrated in Fig. 3. The forward citations of each reference can be collected from the Cited Reference Search in Web of Science (or from other citation databases, such as Google Scholar, Scopus or CiteSeer, see below). The number of references reported by each article is easily calculated once the bibliographic data are downloaded. Then, $\mathbf{A}$, $\mathbf{R}$, $\mathbf{W_1}$ and $\mathbf{W_2}$ can be constructed accordingly (Stage 1). The KHS network and partition process (Stage 2 and Stage 3) can be automatically realized via running an R-program script (available from the contact author).

[13] This is different from the Distinct Author Identification System (DAIS) in Web of Science. Our experiments in DAIS found that the same authors appear in different clusters, and different authors appear in the same cluster. This suggests either transitivity was not adopted in their algorithm (otherwise the same authors should only appear in one author cluster), or limited coverage of publications indexed in WoS (which miss the common links with papers not covered by DAIS). For more discussion on DAIS, see http://science.thomsonreuters.com/support/faq/wok3new/dais/ or the patent application number US20080275859 A1 (Griffith 2008).

Illustrative examples

We test the ASE method on two presumed difficult cases: an American social scientist case with a relatively common name, and a Chinese origin scientist in the nanotechnology domain.

Case 1: John P. Walsh/Walsh, J*

Our first case is "John P. Walsh", which is rendered as "Walsh, J" or "Walsh, JP" in the Web of Science. We start with the "Walsh, J*" case for our first test based on the following three considerations. First, "Walsh, J*" is a relatively popular name. According to the 2000 census Walsh was ranked as the 265th most common surname in the US. Second, we happen to know there are several authors that have publications indexed in Web of Science. And some "Walsh, J" work in a similar or even the same research field. Third, within the examined period (2004–2008) we know at least one "Walsh, J" moved and thus reported different affiliations in his publications. In addition, with one "Walsh, J" being a coauthor of this paper, a cross-checking can be made to further verify the correct classification that we derived from reading each article (which was done independently by the first author). Our goals are twofold: (1) whether this method can correctly identify articles written by this researcher; and (2) how many authors named "Walsh, J*" can be correctly grouped by this approach.

Data processing

The search for published articles written by "Walsh J*" in the last 5 years (2004–2008) was conducted on Feb 19, 2009 in the Social Sciences Citation Index dataset.[14] This returned 125 hits. These full records were exported to VantagePoint data mining software developed by Search Technology, Inc. After removing those articles written by "Walsh, J?" in which "?" is not "P", 69 are left in the database.[15] Among these articles, 72% report cited references, in which 24 shared common references and 26 do not.[16] These 24 aricles associated with "Walsh, J" are published in 17 journals and involved at least 55 authors from 32 research institutions. About 50% of the articles report author full names.[17] The number of reported references ranges from 6 to 162, with a total of 114 unique references appearing at least twice (i.e., are shared by at least two papers). We collected the number of forward citations for each of these 114 references from the Cited References database in

[14] Please note, we use wild-card characters "*" instead of middle initial to relax the formats of reported author names. The "*" is important. Without that, only 47 records were retrieved.

[15] The removal process was conducted in VantagePoint. These 125 records were first clustered into three groups: Group one are articles written by "Walsh, J"; Group two include articles written by "Walsh, J?P"; and Group three consists those written by "Walsh, J?" with"?" not "P". Group 1 and Group 2 are combined which returns 69 articles.

[16] Those records without references are letters, book reviews, etc., which are not the focus of this study.

[17] The full author names are not viewable in the bibliographic data in Web of Science up to September 2006. To get those full names, we came back to the orginal full text of articles. Since different organizations purchased different coverage of full text WoS, in order to do the verification step (but not needed in the ASE algorithm) in some cases we have to look for hard copies or request InterLibrary loans to get them if the electronic version of the full text is not available.

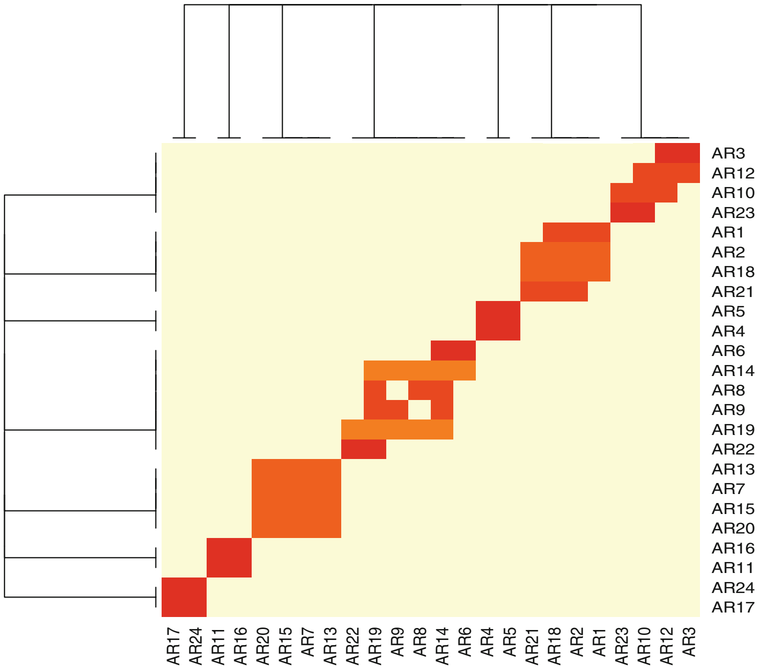

Fig. 4 Heat map on authorship identification

WoS.[18] $\mathbf{W_1}$ and $\mathbf{W_2}$ can then be constructed.[19] We then ran the ASE algorithm on these 24 articles. The other 26 cases will be treated as singletons (unique authors) on the assumption that if they share no references with anyone else in the database, they are unlikely to be any of those authors. We test this assumption below.

Results

An open source software R-program (version 2.8.1) was used to construct the KHS and perform hierarchical clustering with single linkage. Seven structurally equivalent clusters emerged from the corpus of "Walsh, J*" publications. We can use a heat map (Fig. 4) and a dendrogram (Fig. 5) to visualize which articles are structurally equivalent. In the heat map, articles that possess the same color are grouped together. In the dendrogram articles written

[18] In addition to citations received by each reference, we also tried the journal impact factor as the weighting factor. Among the 114 references in the first experiment (Walsh, J*), 23 could not be identified from ISI journal citation reports, suggesting this is less useful than citations for weighting references.

[19] Two weights $\mathbf{W_1}$, $\mathbf{W_2}$ are coded based on the quartile distributions of visibility of references and minimum number of reported references between each pair of targeted papers respectively. For $\mathbf{W_1}$, in both the Walsh, JP and the Li, Y experiments, if a reference was in the first quartile of forward citations, the reference is given a weight of 8, if in the second quartile, the weight is 3, if in the third quartile, the weight is 2 and if in the fourth quartile, the weight is one. For $\mathbf{W_2}$, in the Walsh, JP experiment, if the number of references was in the first quartile of reference counts, the number of references weight is set at 4, if in the second quartile, the weight is 3, if in the third quartile, the weight is 2 and if in the fourth quartile, the weight is one. In the Li, Y experiment, the first quartile of reference counts were given a weight of 8 in $\mathbf{W_2}$. The detailed raw data and coding of KHS variables are available on request.

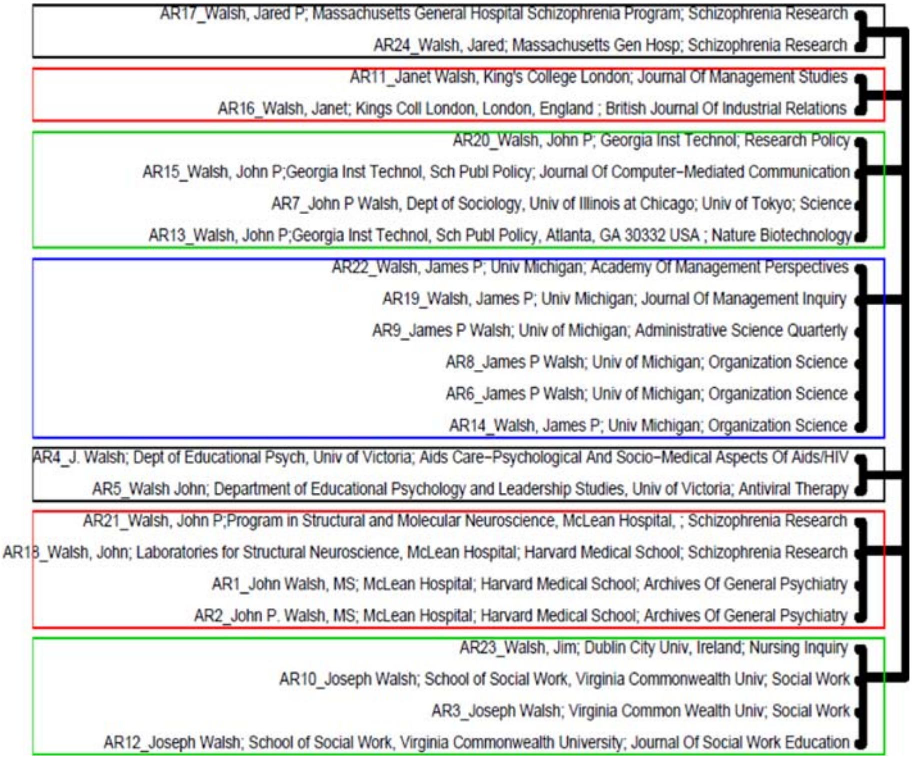
AR17_Walsh, Jared P; Massachusetts General Hospital Schizophrenia Program; Schizophrenia Research
AR24_Walsh, Jared; Massachusetts Gen Hosp; Schizophrenia Research
AR11_Janet Walsh, King's College London; Journal Of Management Studies
AR16_Walsh, Janet; Kings Coll London, London, England ; British Journal Of Industrial Relations
AR20_Walsh, John P; Georgia Inst Technol; Research Policy
AR15_Walsh, John P;Georgia Inst Technol, Sch Publ Policy; Journal Of Computer-Mediated Communication
AR7_John P Walsh, Dept of Sociology, Univ of Illinois at Chicago; Univ of Tokyo; Science
AR13_Walsh, John P;Georgia Inst Technol, Sch Publ Policy, Atlanta, GA 30332 USA ; Nature Biotechnology
AR22_Walsh, James P; Univ Michigan; Academy Of Management Perspectives
AR19_Walsh, James P; Univ Michigan; Journal Of Management Inquiry
AR9_James P Walsh; Univ of Michigan; Administrative Science Quarterly
AR8_James P Walsh; Univ of Michigan; Organization Science
AR6_James P Walsh; Univ of Michigan; Organization Science
AR14_Walsh, James P; Univ Michigan; Organization Science
AR4_J. Walsh; Dept of Educational Psych, Univ of Victoria; Aids Care-Psychological And Socio-Medical Aspects Of Aids/HIV
AR5_Walsh John; Department of Educational Psychology and Leadership Studies, Univ of Victoria; Antiviral Therapy
AR21_Walsh, John P;Program in Structural and Molecular Neuroscience, McLean Hospital, ; Schizophrenia Research
AR18_Walsh, John; Laboratories for Structural Neuroscience, McLean Hospital; Harvard Medical School; Schizophrenia Research
AR1_John Walsh, MS; McLean Hospital; Harvard Medical School; Archives Of General Psychiatry
AR2_John P. Walsh, MS; McLean Hospital; Harvard Medical School; Archives Of General Psychiatry
AR23_Walsh, Jim; Dublin City Univ, Ireland; Nursing Inquiry
AR10_Joseph Walsh; School of Social Work, Virginia Commonwealth Univ; Social Work
AR3_Joseph Walsh; Virginia Common Wealth Univ; Social Work
AR12_Joseph Walsh; School of Social Work, Virginia Commonwealth University; Journal Of Social Work Education

Fig. 5 Cluster dendrogram for “Walsh, J”

by the same author are clustered within one colored frame. These clusters are all crisp clusters in which each element has a clear cut (0/1) membership (Van Mechelen et al. 2004). Compared with the true eight groups which were produced through manual checking, ASE produces only one mis-assignment (AR23), a Type II error of overmatch.[20] We examine this performance more exactly below.

Case 2: “Li, Y”

Asian names, particularly Chinese names, are notoriously challenging for disambiguation (Lin 1988; Tan 1986). Thus, we further test our approach on a more difficult case, a Chinese origin nanoscientist named “Li, Y”. The selection justifications are as follows. To begin with, China has recently become one of the top producers of research papers. Its rapid expanding researcher base, translation & transliteration issues, as well as the existence of overseas Chinese and returnees make it a daunting challenge to distinguish Chinese researchers with the same family name and first initial. We choose nanotechnology in part

[20] A close look tells us the misassignment occurs because AR10, an article reporting few references, shares with AR23 a rarely cited reference. ASE method decides the knowledge homogeneity score between AR10 and AR23 is high enough and thus cluster them together.

because many research evaluations have been conducted in this field. Furthermore, limiting to one field (although a very broad one) increases the difficulty of the problem, since key words, subject categories or other commonly used means of disambiguation may be less effective. We target on authors named "Li, Y" not only because it is one of the most frequent author names appearing in the nano publication database (Table 2),[21] but also because he/she was identified as the most prolific Chinese nanoscientist in a prior study (Kostoff et al. 2006). The use of 2007 is because full author names are viewable on the full record of some journals indexed in WoS since September 2006, making verification easier. A 1-year span of publications also reduces the possibility of researcher mobility which theoretically reduces the error of the "name + affiliation" method, giving us a conservative test of the benefits of the ASE method.

We first extracted all "Li, Y" nano papers published in 2007 from the Georgia Tech nano publication database.[22] This returned 221 hits. This large number of common-named authors in 1 year in one field suggests the magnitude of the problem. After *temporarily* excluding articles that do not share references with the other "Li, Y" articles (treated as singletons), 145 records associated with 376 shared references are eligible for ASE analysis. These articles are published in 82 journals spanning across 33 subject categories as defined by ISI-WoS. About 116 research organizations in 14 countries are involved, and the number of reported references ranges from 7 to 186. Following the same procedure as in the "Walsh, J" case, $\mathbf{A}$, $\mathbf{R}$, $\mathbf{W_1}$ and $\mathbf{W_2}$ are constructed for "Li, Y". The same script was executed and the records were partitioned into 103 predicted clusters, meaning 103 different "Li, Y"s, including many singletons (Fig. 6).[23] In the validation step, using full name and other ancillary data, we find 87 "true" clusters (i.e., unique authors with one or more publications). Twenty-nine records are wrongly assigned, in which 6 cases are Type II error of overmatch, and 23 are Type I error of under match. This leads to an accuracy rate of 80%.[24]

It should be noted that the full names of "Walsh, J" or "Li, Y" and their affiliations in both dendrograms only serve for the reader's convenience. ASE does not need them for authorship identification. It also should be noted that although at first sight the two dendrograms suggest the effectiveness of the "full names + affiliation" method for name disambiguation,[25] these full names and author-matched affiliations are not readily available data in the bibliographic database. Rather, acquiring this information required tremendous hand checking. In addition to insider knowledge, the verification steps of finding out the

[21] Even in the US, Li has become a fairly common name, currently ranked 519th on the list of common names, up from 2084th in the 1990 census.

[22] For a detailed description of this global nano database, please refer to Porter et al. (2008).

[23] Singletons are still possible in the ASE method due to citation weights and matching thresholds rules.

[24] For instance, two articles written by the same author Li Yue at CAS Hefei are wrongly separated due to no shared reference between them. A close examination tells us that these two articles are in different research areas, as seen by their subject category codes. One is in "Chemistry, Physical; Materials Science, Multidisciplinary" and the other is in "Nanoscience & Nanotechnology, Polymer Science". Another example is two articles authored by Li, Ying at Chinese Academy of Sciences (CAS) Shenyang that are in two related but different subject categories. "Acoustics; Chemistry, Multidisciplinary" and "Chemistry, Applied; Engineering, Chemical; Materials Science, Textiles".

[25] This holds if the following three conditions are met: (1) researchers are not mobile; (2) researchers are not affiliated with multiple organizations; and (3) standardized affiliations are reported across different records.

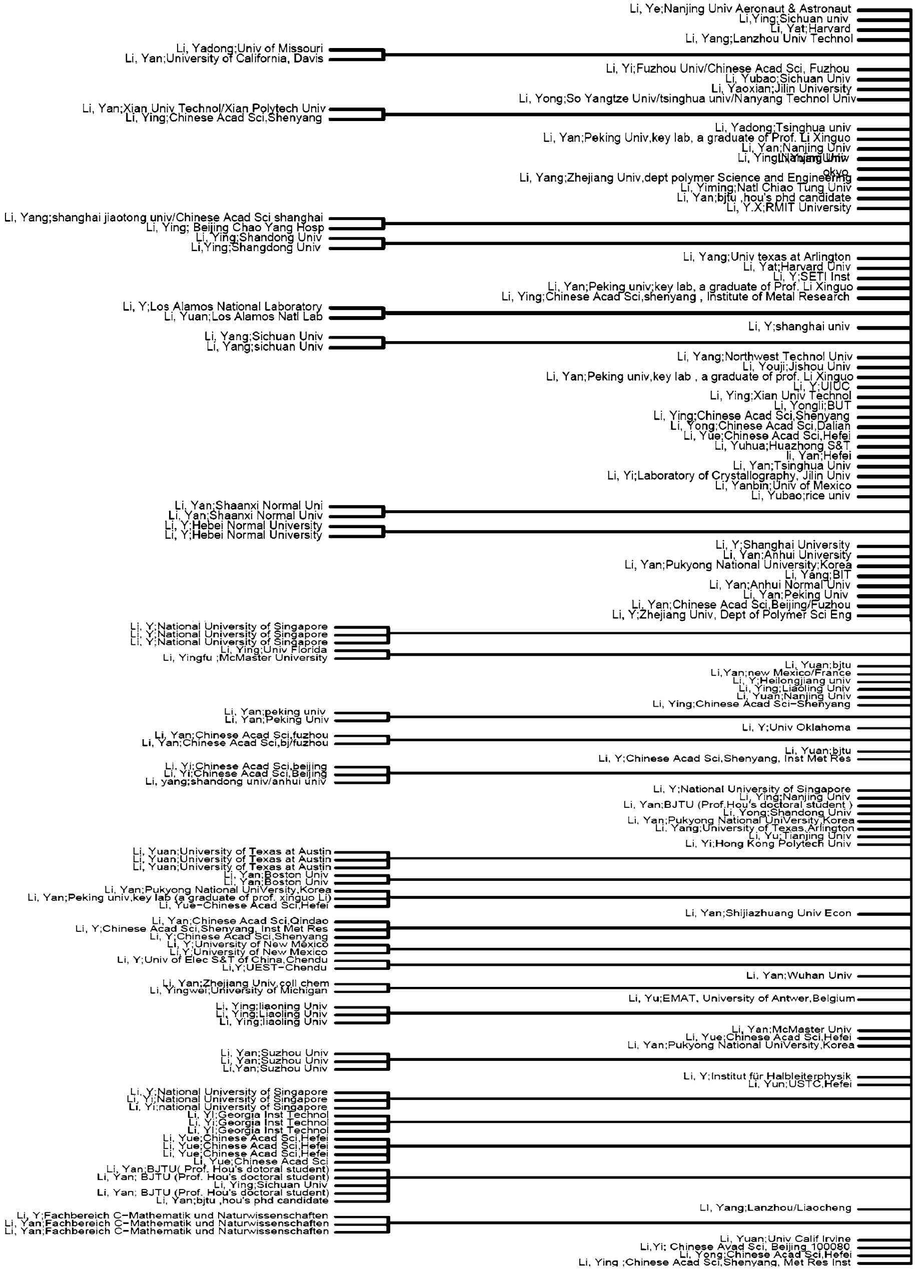
Li, Ye;Nanjing Univ Aeronaut & Astronaut
Li,Ying;Sichuan univ
Li, Yat;Harvard
Li, Yang;Lanzhou Univ Technol
Li, Yadong;Univ of Missouri
Li, Yan;University of California, Davis
Li, Yi;Fuzhou Univ/Chinese Acad Sci, Fuzhou
Li, Yubao;Sichuan Univ
Li, Yaoxian;Jilin University
Li, Yong;So Yangtze Univ/tsinghua univ/Nanyang Technol Univ
Li, Yan;Xian Univ Technol/Xian Polytech Univ
Li, Ying;Chinese Acad Sci,Shenyang
Li, Yadong;Tsinghua univ
Li, Yan;Peking Univ,key lab, a graduate of Prof. Li Xinguo
Li, Yan;Nanjing Univ
Li, Ying;Nanjing Univ
okyo
Li, Yang;Zhejiang Univ,dept polymer Science and Engineering
Li, Yiming;Natl Chiao Tung Univ
Li, Yan;bjtu ,hou's phd candidate
Li, Y.X;RMIT University
Li, Yang;shanghai jiaotong univ/Chinese Acad Sci shanghai
Li, Ying; Beijing Chao Yang Hosp
Li, Ying;Shandong Univ
Li,Ying;Shangdong Univ
Li, Yang;Univ texas at Arlington
Li, Yat;Harvard Univ
Li, Y;SETI Inst
Li, Yan;Peking univ;key lab, a graduate of Prof. Li Xinguo
Li, Ying;Chinese Acad Sci,shenyang , Institute of Metal Research
Li, Y;Los Alamos National Laboratory
Li, Yuan;Los Alamos Natl Lab
Li, Y;shanghai univ
Li, Yang;Sichuan Univ
Li, Yang;sichuan Univ
Li, Yang;Northwest Technol Univ
Li, Youji;Jishou Univ
Li, Yan;Peking univ,key lab , a graduate of prof. Li Xinguo
Li, Y;UIUC
Li, Ying;Xian Univ Technol
Li, Yongli;BUT
Li, Ying;Chinese Acad Sci,Shenyang
Li, Yong;Chinese Acad Sci,Dalian
Li, Yue;Chinese Acad Sci,Hefei
Li, Yuhua;Huazhong S&T
li, Yan;Hefei
Li, Yan;Tsinghua Univ
Li, Yi;Laboratory of Crystallography, Jilin Univ
Li, Yanbin;Univ of Mexico
Li, Yubao;rice univ
Li, Yan;Shaanxi Normal Uni
Li, Yan;Shaanxi Normal Univ
Li, Y;Hebei Normal University
Li, Y;Hebei Normal University
Li, Y;Shanghai University
Li, Yan;Anhui University
Li, Yan;Pukyong National University;Korea
Li, Yang;BIT
Li, Yan;Anhui Normal Univ
Li, Yan;Peking Univ
Li, Yan;Chinese Acad Sci,Beijing/Fuzhou
Li, Y;Zhejiang Univ, Dept of Polymer Sci Eng
Li, Y;National University of Singapore
Li, Y;National University of Singapore
Li, Y;National University of Singapore
Li, Ying;Univ Florida
Li, Yingfu ;McMaster University
Li, Yuan;bjtu
Li,Yan;new Mexico/France
Li, Y;Heilongjiang univ
Li, Ying;Liaoling Univ
Li, Yuan;Nanjing Univ
Li, Ying;Chinese Acad Sci−Shenyang
Li, Yan;peking univ
Li, Yan;Peking Univ
Li, Y;Univ Oklahoma
Li, Yan;Chinese Acad Sci,fuzhou
Li, Yan;Chinese Acad Sci,bj/fuzhou
Li, Yuan;bjtu
Li, Y;Chinese Acad Sci,Shenyang, Inst Met Res
Li, Yi;Chinese Acad Sci,beijing
Li, Yi;Chinese Acad Sci,Beijing
Li, yang;shandong univ/anhui univ
Li, Y;National University of Singapore
Li, Ying;Nanjing Univ
Li, Yan;BJTU (Prof.Hou's doctoral student)
Li, Yong;Shandong Univ
Li, Yan;Pukyong National UniVersity,Korea
Li, Yang;University of Texas,Arlington
Li, Yu;Tianjing Univ
Li, Yi;Hong Kong Polytech Univ
Li, Yuan;University of Texas at Austin
Li, Yuan;University of Texas at Austin
Li, Yuan;University of Texas at Austin
Li, Yan;Boston Univ
Li, Yan;Boston Univ
Li, Yan;Pukyong National UniVersity,Korea
Li, Yan;Peking univ,key lab (a graduate of prof. xinguo Li)
Li, Yue−Chinese Acad Sci,Hefei
Li, Yan;Shijiazhuang Univ Econ
Li, Yan;Chinese Acad Sci,Qindao
Li, Y;Chinese Acad Sci,Shenyang, Inst Met Res
Li, Y;Chinese Acad Sci,Shenyang
Li, Y;University of New Mexico
Li,Y;University of New Mexico
Li, Y;Univ of Elec S&T of China,Chendu
Li,Y;UEST−Chendu
Li, Yan;Wuhan Univ
Li, Yan;Zhejiang Univ,coll chem
Li, Yingwei;University of Michigan
Li, Yu;EMAT, University of Antwer,Belgium
Li, Ying;liaoning Univ
Li, Ying;Liaoling Univ
Li, Ying;liaoling Univ
Li, Yan;McMaster Univ
Li, Yue;Chinese Acad Sci,Hefei
Li, Yan;Pukyong National UniVersity,Korea
Li, Yan;Suzhou Univ
Li, Yan;Suzhou Univ
Li,Yan;Suzhou Univ
Li, Y;Institut für Halbleiterphysik
Li, Yun;USTC,Hefei
Li, Y;National University of Singapore
Li, Yi;National University of Singapore
Li, Yi;national University of Singapore
Li, YI;Georgia Inst Technol
Li, Yi;Georgia Inst Technol
Li, YI;Georgia Inst Technol
Li, Yue;Chinese Acad Sci,Hefei
Li, Yue;Chinese Acad Sci,Hefei
Li, Yue;Chinese Acad Sci,Hefei
Li, Yue;Chinese Acad Sci
Li, Yan;BJTU(Prof. Hou's dotoral student)
Li, Yan; BJTU (Prof. Hou's doctoral student)
Li, Ying;Sichuan Univ
Li, Yan; BJTU (Prof. Hou's doctoral student)
Li, Yan;bjtu ,hou's phd candidate
LI, Yang;Lanzhou/Liaocheng
Li, Y;Fachbereich C−Mathematik und Naturwissenschaften
Li, Yan;Fachbereich C−Mathematik und Naturwissenschaften
Li, Yan;Fachbereich C−Mathematik und Naturwissenschaften
Li, Yuan;Univ Calif Irvine
Li,Yi; Chinese Avad Sci, Beijing 100080
Li, Yong;Chinese Acad Sci,Hefei
Li, Ying ;Chinese Acad Sci,Shenyang, Met Res Inst

Fig. 6 Dendrogram of "Li, Y" case

full names and affiliations of "Li, Y" from the original papers is time consuming (see below). In some cases, we looked for author names in Chinese characters via checking Chinese journal websites and online CVs if they are available.

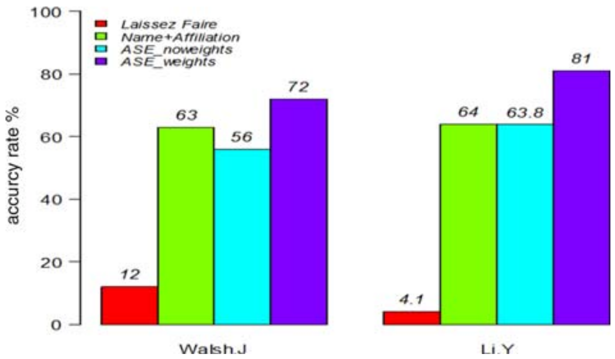

Fig. 7 Comparison of name disambiguation methods (as discussed earlier, the third category of extant disambiguation methods require comprehensive data collection and coding, and involve lots of subjective decisions when dealing with missing data. We could not replicate them in these two cases and thus have not included these more elaborate methods for accuracy rates comparison. In terms of time spent, it is reasonable to believe they are more time consuming than the ASE or name + affiliation methods)

Benchmarking and comparison

So how did the ASE method perform in the two cases? We evaluate its effectiveness on two dimensions: correct classification rate and time spent.

Three commonly used methods which can be replicated are benchmarked: simply grouping (laissez faire), name + affiliation, and ASE without citation weights. In order to avoid underestimating the effectiveness of the other methods, we purposely show their optimal results if subjective choices are needed. Take simply name matching for example: the number of the largest group is taken as the correctly assigned records. In the case of "Walsh, J", the largest true cluster has six papers, so an accuracy of 25%, i.e., 6/24 is taken for the approach of simply name matching. Accuracy would be lower if we chose a random author or the average accuracy as the benchmark. For the method name + affiliation, 100% match between reported name and affiliation in the publication are assumed for those records if the targeted researcher is the only author, or reprint author, or one single affiliation reported; and a 50% accuracy rate, our best guess, is assumed for records which do not fit into any of the above three situations. In the "Walsh, J" case, 13 out of 24 records report only one affiliation or report "Walsh, J*" as the sole author or the reprint author. Within these 24 articles, "Walsh, John P" is involved in six articles reporting three different institutions: Georgia Tech (three times), University of Illinois at Chicago (twice), and University Tokyo (once), thus yielding at least three mistakes. Accordingly, the highest accuracy rate of this method, even assuming no typos and translation problems, is 65%.[26]

In the case of "Li, Y", 55 articles report one affiliation, and 27 have "Li, Y" as the reprint author, which adds up to 72 papers identifiable with Li, Y and his/her affiliation after removing overlapping articles between the two conditions. Again assuming no mis-assignment among these 72 papers, which involved lots of efforts of manual

[26] I.e., ((13 − 3) + (24 − 13) * 0.5)/24 * 100%. The mis-assignments (two in Illinois at Chicago and one in University of Tokyo), are within these 13 records with identifiable affiliation.

standardization and cross checking affiliation names, the highest accuracy rate would be 63%.[27]

Recall in both the "Walsh, JP" and "Li, Y" cases, a large proportion of records—26 out of 50 for "Walsh, JP" and 76 out of 221 for "Li, Y"—were not included in the ASE clustering due to no overlapping references between them and the others. Excluding them would severely limit the applicability of ASE. So to evaluate the usefulness of the ASE method, we need to benchmark accuracy rates with all records included. ASE analysis will automatically take those records without shared references as singletons. For instance, "Walsh, JP" associated with those 26 records will be regarded as 26 different authors. In the same vein, 76 authors named "Li, Y" will be classified as different researchers from any "Li, Y" associated with the rest of the 145 records. We then check these by hand to see how these singletons affect the accuracy rate.[28] Figure 7 compares the accuracy rates among the laissez faire method, name + affiliation method, ASE without weighting, and ASE with weighting method (including singletons in each case).

First, the ASE approach (with weighting) produces the highest accuracy rates, followed by name + affiliation, while simple grouping yields the lowest rate. As expected, the simple grouping method performs even poorer in the Chinese author name case given all the problems discussed before. Their low correct classification rates confirm that name ambiguity is an important problem and has to be dealt with before conducting any rigorous bibliographical analysis at the individual level. Surprisingly, the ASE approach still outperforms others in the single field (nanotechnology) case, where we might expect more false positives due to common field references (and a higher performance for the name + affiliation benchmark). This suggests that the ASE method may be especially powerful in exactly those cases where other methods based on common field keywords, etc. may have the most difficulty. We also see the importance of the weighting scheme. Weighting the algorithm by (the inverse of) the citation frequency of references results in a significant improvement in accuracy, about a 25% increase for both cases.[29]

Given the importance of the weightings, we also test the ASE method using other sources for the forward citations data: Google Scholar and Scopus, which have become increasingly popular as alternatives or complements to WoS.[30] For the case of "Walsh, J", two experiments have been done. One is to test whether different citation sources (Google Scholar or Scopus) would influence the partition results or not. We retrieved the forward citations of these shared 114 references from both sources,[31] and reran the R script. We find that the ln(cites) in WoS and Google Scholar are correlated .83, while ln(cites) in Scopus are correlated with the counts from the other two databases in the range of .35–.46. The results show ASE method is rather robust across these three citation databases, which is not surprising, given the high correlations among the citation counts recovered from the

[27] The formula is calculated by $((72 - 17) + (145 - 72) * 0.5))/145 * 100$, where 17 is the number who are wrongly assigned due to different authors possessing exactly the same English translated name in the same organization.

[28] It turns out a large proportion of records (50% in the Walsh, J case and 83% in Li, Y) that do not share citations are in fact singletons (unique authors).

[29] The differences that weighting makes are even larger if only records with shared citations considered.

[30] For more details on the comparative advantages of these databases, please refer to Meho and Yang (2007), Pauly and Stergiou (2005), and Zhao and Logan (2002).

[31] All records downloading and data from Google Scholar and Scopus were completed November 25–30, 2009. These new measures of forward citations change the weighting of $\mathbf{W_1}$ and therefore the knowledge homogeneity scores used to calculate the clustering.

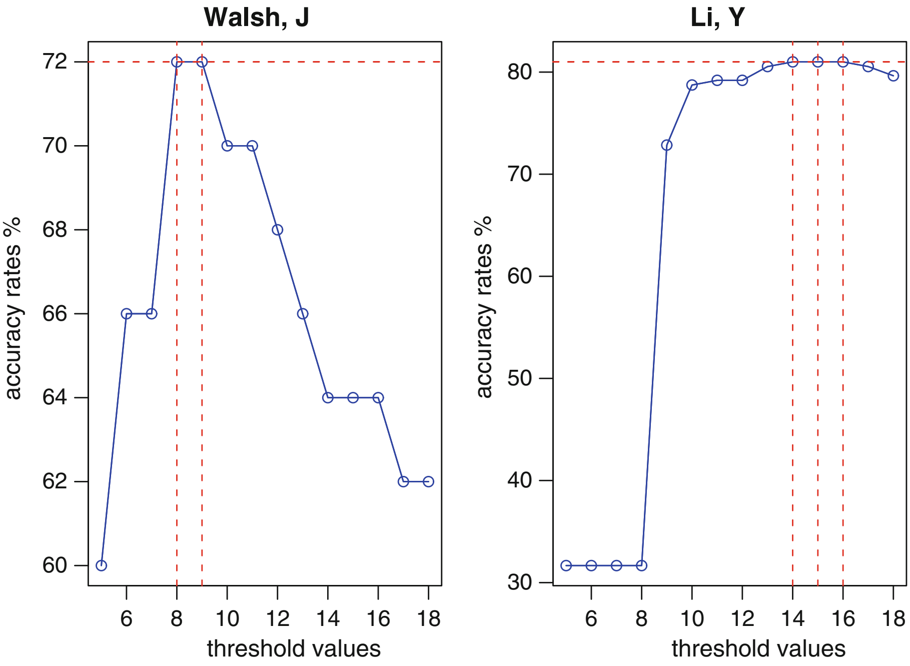

Fig. 8 Sensitivity tests on knowledge homogeneity thresholds

different databases.[32] As a second robustness check, we repeat the whole process of "Walsh, J" name disambiguation in Scopus, and our ASE method produces a 100% accuracy rate for the 17 articles sharing at least one reference.[33] Thus, this method seems fairly robust to difference sources of citations for the weighting matrix ($\mathbf{W_1}$) and to the use of different bibliometric databases (WoS or Scopus). Of course, given the full names and author-affiliation information in Scopus, name + affiliation will also produce higher results using Scopus (94% accuracy in this case).

Besides the weighting mechanism, another key component in our algorithm is the threshold for the knowledge homogeneity score, the minimum value of summed weighted common references that two papers need to have before they are considered in the same cluster. These values were selected based on both intuition (i.e., either one rarely cited reference or several common references) and several rounds of trial and error. As a robustness check, we changed the KHS threshold values (sum of weighted shared references) in both cases and re-estimated the accuracy rates for the resulting ASE clustering. The results are demonstrated in Fig. 8. In the case of Walsh, J*, the accuracy score would

[32] One record in Google Scholar and three records in Scopus are mis-assigned compared to one mis-assignment in WoS. A closer examination indicates the messier formats of references in Scopus and inconsistent coverage of journals partially accounts for Scopus having a lower correlation in citations with WoS and Google Scholar. The detailed analyses and results are available on request.

[33] To make the search comparable to the "Walsh, J" case indexed in WoS-SSCI, the search strategy in Scopus is also confined to social science using the query of "*AUTHOR-NAME(walsh, j*) AND SUBJAREA (mult OR arts OR busi OR deci OR econ OR psyc OR soci) AND (PUBYEAR BEF 2009) AND (PUBYEAR AFT 2003)*". This returned 164 hits. After removing "Walsh, J?" where "?" is not P, 41 records were left, 37 of them report references, and 17 records share at least one reference with the others.

be greater than 60% if the KHS threshold is located anywhere between 5 and 18. And the accuracy rate is no less than 70% if the value is between 8 and 11. For the large dataset from Li, Y, any threshold between 13 and 18 yields an accuracy score around 80%. These results show that the accuracy rates are not very sensitive to the threshold values. This is particularly true in the case of the nanotechnology scientists. It is also interesting to note that the "turning points" for the accuracy rates show a different pattern in the domains of social sciences and nanotechnology. One possible explanation would be researchers in the same or similar domain are more likely to cite similar references, which requires a higher threshold to distinguish differences among researchers.

In addition to the accuracy rate, we also evaluate the resource requirements of these methods. In the example of "Li, Y" with 221 records, a total of 6 h were used to complete the ASE method from scratch, including downloading data from WoS, constructing the authorship-citation matrix, weights and executing analysis.[34] By contrast, matching "Li, Y" with his/her reported affiliation (name + affiliation) took over 15 h and the "truth" group, which was produced by manual checking, took over 23 h due to the factors discussed before. Thus, not only is the ASE method more accurate, it is also less time consuming than name + affiliation.[35]

Discussion

Authorship identification is a pervasive challenge for current bibliometrics analyses and research evaluation. This article proposes an ASE approach to the name ambiguity problem, based on the attributes of common references (or lack thereof). The method proposed here is limited in the following ways. First, the underlying notion of this method is that each individual's fixed knowledge stock in a given time period makes his/her reference coverage different from the others. This is the fundamental assumption to differentiate articles reporting the same family name and first initial. Put another way, the two assumptions distinguishing intra-author variation from inter-author variation are as follows:

1. Any articles written by the same author has at least one *single* (direct or indirect) linkage with the rest of his articles;
2. The knowledge homogeneity score of articles written by different authors, regardless of their name overlap, is below a latent threshold.

If assumption 1 is violated, i.e., the author has broad research interests or he shifts his research agenda over time such that two papers he authored have no shared citations at all (directly or through third articles), his shared authorship will not be captured by this approach. In fields where the number of citations allowed per article is smaller, this problem is greater. If assumption 2 does not hold, i.e., two researchers with the same family name and first initial work in the same field and read largely overlapping literature, they may be taken as the same author by this method. Our experiments show this assumption largely holds, and that, in the large majority of cases, an ASE with weighting

[34] We did not make a memo of time spent in the case of "Walsh, J" because of the testing and revising efforts in this first prototype.

[35] Some text mining software (such as VantagePoint) has installed a name + affiliation function. We ran it in VantagePoint using person name fuzzy matching and verified by organization name matching in the case of Li, Y, and the performance is very poor and, in spite of the automation, took significant time.

algorithm will produce accurate partitioning of a set of authors. This suggests that the cognitive mapping assumption underlying this method holds. Researchers draw on a unique set of references in their work (with rare references being especially individualized) and their publications reflect their knowledge set.

The main limitation of ASE is its reliance on the availability of data on forward citations to references (for constructing the $\mathbf{W_1}$ matrix). In both cases, 85% of the time spent on the ASE method is downloading citation counts for the references. If this process can be automated in the future, ASE will become much more efficient. Additionally, the sensitivity test suggests the knowledge homogeneity threshold should be set sufficiently high, and may vary in different disciplines. Future work might use this fact as a means of exploring the underlying cognitive maps of different fields. In this paper we test this method against two sets of predefined articles given their presumed difficulties. Its performance in a general setting needs further exploration.

One factor that may muddy the bibliometric fingerprints is the existence of collaborations and references introduced by other coauthors. A previous study shows that co-authorship can serve as a good algorithm dealing with homonyms issue (Wooding et al. 2005), and co-authors are used as a component in many of the more sophisticated algorithms, but not in the ASE method. Theoretically, the impact of co-authorship on our ASE approach is mixed. On the one hand, references brought by different authors might muddy the fingerprint. On the other hand, new literature brought in by co-authors is likely to be added to the target author's cognitive map and hence reappear in future papers by that author (even if these future papers do not have the same co-authors), leading the ASE method to recognize these as sharing an author. In our experiments, we find that ASE accuracy rate is unrelated to the number of authors on the papers (results available from contact author).

Thus, in spite of the above limitations, this method provides a promising means of authorship disambiguation at a large-scale. Compared with other methods, ASE provides the following advantages. First, this approach yielded more accurate results than common methods, especially for the Chinese name, when it is extremely difficult to distinguish authorship even based on their full English names. This is also related to its second advantage, theoretical independence of spellings of author and affiliation names (since the algorithm matches authors based on references). So potentially this method is applicable for databases regardless of author names. Thirdly, the ASE approach is less time consuming than name + affiliation method and other statistical methods. Once the coding mechanism and threshold value are set, the same R script can run on different cases without modification. In addition, different from other methods, which performed better in small databases, the ASE method would likely be even more useful in the case of large datasets due to automatic clustering and a larger set of shared references. In addition, this method may be more accurate for tracing mobile authors or inventors, since it does not use affiliations or collaborators as part of the identification strategy (cf. Trajtenberg et al. 2006; Raffo and Lhuillery 2009).

The paper contributes to the name disambiguation literature in two aspects. First, it provides new evidence that who is who is not a trivial problem, and any *rigorous* analysis at the individual level has to deal with that to make a convincing result. Second, researchers' latent knowledge scope alone is a good discriminator for name disambiguation, and examining shared references can be an effective way to trace the bibliometric fingerprints of authors. We agree with MacRoberts' statement that "identification of individual authors or institutions cannot be accomplished solely by computer analysis" (MacRoberts and MacRoberts 1989), but given the limitations and advantages of the ASE method, we also believe it is a good *alternative* when manual checking is too costly or impossible. In this study, ASE is applied on two sets of publications derived from the ISI-WoS database and replicated for one of them using Scopus (as well as using Google Scholar and Scopus as alternative sources of weights for the WoS case). In the future, we plan to improve our method in three directions. First, relax this assumption of similar author name spelling and test its performance again. Second, examine the sensitivity of weight coding and thresholds on ASE performance. Third, test it on other large scale archival data such as patents. And, finally, we plan to investigate inventor mobility and collaboration networks using a disambiguated dataset of patents or papers.

Acknowledgments The authors would like to thank Dr. Diana Hicks and participants in the Workshop for Original Policy Research at Georgia Tech, and the Research Institute of Economics, Technology and Industry, Tokyo, seminar series for their comments on earlier draft, Dr. Juan D. Rogers for offering the R-class and useful discussion, Dr. Philip Shapira for the use of the global nano publication data in the study, and the anonymous reviewer for his useful suggestions. We also would like to thank authors and co-authors of "Li, Y" for clarifying their publication lists.

References

Abbasi, A., & Chun, H. (2006). Visualization authorship for identification. In: S. Mehrotra, et al. (Eds.), *Proceedings of the IEEE international conference on intelligence and security informatics (LNCS 3975)* (pp. 60–71). Berlin: Springer-Verlag.

Borgman, C. L., & Siegfried, S. L. (1999). Getty's Synoname^{TM} and its cousins: A survey of applications of personal name-matching algorithms. *Journal of the American Society for Information Science, 43*(7), 45–476.

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence, 4*(1), 1–13.

Frietsch, R., Tang, L., & Hinze, S. (2008). *Bibliometric data study: Assessing the current ranking of the People's Republic of China in a set of research fields*. Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis, No 15. Karlsruhe: Fraunhofer ISI.

Garfield, E. (1969). British quest for uniqueness versus American egocentrism. *Nature, 223*(5207), 763.

Griffith, R. A. (2008). Method and system for disambiguating informational objects. Patent Application Number: US 20080275859 A1. USPTO.

Han, H., Giles, C. L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two supervised learning approaches for name disambiguation in author citations*. Paper presented at the Proceedings of the ACM/IEEE Joint Conference on Digital Libraries.

Han, H., Xu, W., Zha, H., & Giles, C. L. (2005). *A hierarchical naive Bayes mixture model for name disambiguation in author citations*. Paper presented at the Proceedings of the 2005 ACM Symposium on Applied Computing.

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In M. Marlino, T. Sumner, & F. M. Shipman III (Eds.), *Proceedings of the 5th ACM/IEEE joint conference on digital libraries* (pp. 334–343). Denver: ACM Press.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside. http://faculty.ucr.edu/~hanneman/.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In J. Euzenat & J. Domingue (Eds.), *Proceedings of the 12th international conference on artificial intelligence: Methodology, systems, applications (AIMSA'06), LNCS 4183* (pp. 77–86). Berlin: Springer-Verlag.

Huang, J., Ertekin, S., & Giles, C. L. (2006). *Fast author name disambiguation in CiteSeer*. Working paper. http://www.cse.psu.edu/~sertekin/Papers/IST-TR_DisambiguationCiteseer.pdf.

Jacobs, L. F., & Schenk, F. (2003). Unpacking the cognitive map: The parallel map theory of hippocampal function. *Psychological Review, 110*(2), 285–315.

Jones, B., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science, 322*, 1259–1262.

Kang, I. S. (2009). On co-authorship for author disambiguation. *Information Processing and Management, 45*(1), 84–97.

Kostoff, R. (2008). Comparison of China/USA science and technology performance. *Journal of Informetrics, 57*, 1–10.

Kostoff, R., et al. (2006). *The structure and infrastructure of Chinese science and technology*. DTIC Technical Report, No. ADA 443315. http://www.onr.navy.mil/sci_tech/33/332/docs/060307_chinese_sci_tech.pdf.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lai, R., D'amour, A., & Fleming, L. (2009). *The careers and co-authorship networks of U.S. patent-holders since 1975*. Working paper.

Lin, J. C. (1988). Chinese names containing non-Chinese given name. *Cataloging & Classification Quarterly, 9*(1), 69–81.

Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology, 1*, 49–80.

Macroberts, M. H., & Macroberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*, 342–349.

McCallum, A., & Wellner, B. (2003). *Toward conditional models of identity uncertainty with application to proper noun conference*. Paper presented at the IJCAI Workshop on Information Integration.

Meho, L., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus Scopus and Google Scholar. *Journal of the American Society for Information Science & Technology, 58*(13), 2105–2125.

Merton, R. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

National Science Foundation (NSF). (2008). *Science and engineering indicators*. Washington: Government Printing Office.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, England: Oxford University Press.

Pasula, H., Marthi, B., Milch, B., Russell, S., & Shpitser, I. (2004). *Identity uncertainty and citation matching*. Paper presented at the Advances in Neural Information Processing (NIPS).

Pauly, D., & Stergiou, K. I. (2005). Equivalence of results from two citation analyses: Thomson ISI's citation index and Google's scholar service. *Ethics in Science and Environmental Politics, 5*, 33–35.

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

Pieters, R., Baumgartner, H., Vermunt, J., & Bijmolt, T. (1999). Importance and similarity in the evolving citation network of the International Journal of Research in Marketing. *International Journal of Research in Marketing, 16*(2), 113–127.

Porter, A. L., Youtie, J., Shapira, P., & Schoneneck, D. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research, 10*, 715–728.

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

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. In B. Cronin (Ed.), *Annual review of information science and technology* (Vol. 43). Maryland, USA: American Society for Information Science and Technology (ASIST).

Soler, J. M. (2007). Separating the articles of authors with the same name. *Scientometrics, 72*(2), 281–290.

Strotmann, A., Zhao, D., & Bubela, T. (2009). *Author name disambiguation for collaboration network analysis*. Working paper.

Tan, C. N. (1986). Chinese personal names. *Library Association Record, 88*, 551.

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD), 3*(3), Article 11.

Trajtenberg, M., Shiff, G., & Melamed, R. (2006). *The "names game": Harnessing inventors' patent data for economic research*. NBER Working Paper No. 12479.

Treeratpituk, P., & Giles, C. L. (2009). *Disambiguating authors in academic publications using random forests*. Paper presented at the JCDL, Austin, Texas, USA.

Van Mechelen, I., Bock, H. H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research, 13*(5), 363–394.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Wooding, S., Wilcox-Jay, K., Lewison, G., & Grant, J. (2005). Co-author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. *Scientometrics, 66*(1), 11–21.

Youtie, J., Shapira, P., & Porter, A. (2008). National nanotechnology publications and citations. *Journal of Nanoparticle Research, 10*(6), 981–986.

Zhao, D. Z., & Logan, E. (2002). Citation analysis using scientific publications on the web as data source: A case study in the XML research area. *Scientometrics, 54*(3), 449–472.

Zhou, P., & Leydesdorff, L. (2008). China ranks second in scientific publications since 2006. *ISSI Newsletter, 13*, 7–9.