Scientometrics (2013) 96:683–697
DOI 10.1007/s11192-013-0978-8

Author name disambiguation in scientific collaboration and mobility cases

Jiang Wu · Xiu-Hao Ding

Received: 4 November 2012 / Published online: 24 February 2013

Abstract Scientists generally do scientific collaborations with one another and sometimes change their affiliations, which leads to scientific mobility. This paper proposes a recursive reinforced name disambiguation method that integrates both coauthorship and affiliation information, especially in cases of scientific collaboration and mobility. The proposed method is evaluated using the dataset from the Thomson Reuters Scientific "Web of Science". The probability of recall and precision of the algorithm are then analyzed. To understand the effect of the name ambiguation on the *h*-index and *g*-index before and after the name disambiguation, calculations of their distribution are also presented. Evaluation experiments show that using only the affiliation information in the name disambiguation achieves better performance than that using only the coauthorship information; however, our proposed method that integrates both the coauthorship and affiliation information can control the bias in the name ambiguation to a higher extent.

Keywords Author disambiguation · Scientific collaboration · Scientific mobility · Coauthorship · Affiliation

Introduction

Name disambiguation remains one of the various challenges in bibliometrics and the major obstacle in studies being performed in many disciplines (Smalheiser and Torvik 2009). The basic issue of name disambiguation is on how to distinguish the papers of an author from all the other papers written by his/her namesakes (Onodera et al. 2011). In other words, given a number of papers written by a namesake, the papers that belong to several different authors with a particular namesake are clustered separately according to the distance

J. Wu
School of Information Management, Wuhan University, Wuhan 430072, China
e-mail: jiangwu.john@gmail.com

X.-H. Ding (✉)
School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: dingxiuhao@gmail.com

between pairwise papers (Huang et al. 2006; Soler 2007). Because of the limitation caused by indicating the authors only by their last name and the initials of their first and middle names in the Thomson Reuters Scientific (ISI) Web of Science (WoS), the author name might be related to several different authors, which generates the so-called author homonym problem in name ambiguation (Kang et al. 2009). Thoroughly solving this problem is impossible because data such as affiliations, e-mails, coauthors, references, and personal webpages are required, which are difficult to access and integrate in practice (Huang et al. 2006; Onodera et al. 2011; Soler 2007). Although Gurney et al. (2012) has recently attempted to merge multi-information to calculate the multi-aspect similarity of name disambiguation, moving from the proof-of-concept to the working process still requires much effort, especially in large-scale databases. Among these data, coauthorship and affiliation information are the most important factors in name disambiguation.

Coauthorship is the easiest way to access (and has been regarded as the most distinguishable feature) in the name disambiguation (Kang et al. 2009; Wooding et al. 2006), which is based on the assumption that the identity of an author is characterized by his/her coauthors (Kang et al. 2009). The algorithm is usually recursively executed. Starting from a paper by the searched author, papers with at least one common coauthor that shared in that particular paper are then classified into a cluster. The list of coauthors of the searched author is updated during the recursive process (Wooding et al. 2006). Unfortunately, the use of coauthorship fails when two papers do not share common coauthors although they might be indeed written by the same author. This disadvantage could be because scientists collaborate with completely different people to widen the scope of their research topics, or they move to a new institution and find new collaborators (this situation usually happens after students finish their Ph.D. at a certain university and then find a faculty position in another university). Furthermore, single-author papers also cause problems that influence the accuracy of this algorithm because of the lack of coauthorship information. The above problems could lead to incorrect omission of the papers (*false negatives, lack of recall*). Moreover, cases occur where a common name of the coauthors is found because the common name may correspond to different authors, and the papers that share that particular common coauthor name could possibly be written by different authors who collaborate with some other different authors. This problem leads to the incorrect assignment of the papers (*lack of precision*).

As an example, we consider the papers of Christopher C. Yang, who has two affiliations—Chinese University Hong Kong and Drexel University, as shown in Fig. 1. Based on the coauthorship information only, identifying these papers if they are written by the same "Yang, CC" is impossible because the coauthors in these papers are completely different. Based on *the diversity of collaborations,* the affiliation information could aid in the disambiguation of this author name.

2009|YANG, CC;LIU, N|*Drexel Univ*|*Drexel Univ*;Chinese Univ Hong Kong|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY|000263935100006

2002|YANG, CC;CHUNG, A|*Chinese Univ Hong Kong*|*Chinese Univ Hong Kong*|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY|000173320900011

2000|YANG, CC;LUK, JWK;YUNG, SK;YEN, J|*Chinese Univ Hong Kong*|*Chinese Univ Hong Kong*|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE|000085224800004

Fig. 1 Publication samples of "Yang, CC"

The affiliation information of authors is regarded as another important feature in disambiguating the author names. However, the rapid growth in the number of active scientists and the emergence of interdisciplinary collaborations make the affiliation information insufficient in assisting in the author disambiguation (Tang and Walsh 2010). In the ISI WoS dataset, for papers with multi-authors and multi-affiliations, the affiliation and authorship do not always correctly match (Tang and Walsh 2010). Furthermore, authors with the same names might belong to the same affiliation as well. Distinguishing them using only their affiliation information is impossible. Moreover, in some publications, the same affiliation may be identified by the authors by different names, e.g., "Chinese Academy of Sciences" is usually abbreviated as "Academia Sinica," "CAS," and "CHINESE ACAD SCI". Further, "Eidgenössische Technische Hochschule Zürich" is sometimes abbreviated as "ETH Zurich," and "Peking University" is sometimes used as "Beijing University." Some authors may also name differently their affiliations (including the abbreviation) in their publications during their different career periods. In addition, the name of a university might be changed. For example, Zhongshan Medical University changed its name to Zhongshan University after being merged in 2001, which is also commonly referred to as Sun Yat-sen University. No uniform standard exists to define the affiliation information; therefore, using only the affiliation information could lead to a low probability of recall and precision (Huang et al. 2006).

Figure 2 shows the published papers of Loet Leydesdorff that used completely different affiliations—Univ. Amsterdam and Dept Sci & Technol Dynam. Name disambiguation by calculating the similarity of these two affiliations is impossible (Onodera et al. 2011). Therefore, using only the affiliation information is insufficient to identify the author. Using the coauthorship information might build a relationship between two different affiliations and provide a solution for the name disambiguation in event of *the movement of the scientists*.

Therefore, considering separately the coauthorship and affiliation information is not enough to disambiguate the author names. In this paper, we propose a recursive method that combines the coauthorship and affiliation information to disambiguate the author names in the events of *scientific collaboration and mobility*. In this method, the coauthorship and affiliation information reinforce each other to increase the recall probability and precision of the name disambiguation. In "Methods" section, the methods are introduced. In "Evaluation experiments" section, the evaluation experiments based on the dataset retrieved from the ISI WoS are explained. In "Discussions and conclusion" section, the discussions and conclusion are provided.

Methods

The proposed method integrates the coauthorship and affiliation information to disambiguate the author names. Let us assume that two papers are written by the same author,

1990|LEYDESDORFF, L|*UNIV AMSTERDAM*| |SCIENTOMETRICS|A1990CY00700008

1996|*VANDENBESSELAAR, P*; LEYDESDORFF, L|*UNIV AMSTERDAM*|UNIV AMSTERDAM|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE|A1996UM12100003

1997|LEYDESDORFF, L;*VANDENBESSELAAR, P*|*DEPT SCI & TECHNOL DYNAM*|*DEPT SOCIAL SCI INFORMAT*|SCIENTOMETRICS|A1997WP65300011

1994|LEYDESDORFF, L|*DEPT SCI & TECHNOL DYNAM*| |SCIENTOMETRICS|A1994PD56500004

Fig. 2 Publication samples of "Leydesdorff, L"

which share a namesake with another author; these papers have at least one common coauthor, and the authors with namesakes belong to the same affiliation and are engaged in a particular scientific field. Suppose that $P = \{p_1, p_2, \ldots, p_n\}$ is the set of papers written by the author with a namesake S; then $a_1, a_2, \ldots, a_n$ are the corresponding affiliations of the author in these papers, and $c_1, c_2, \ldots, c_n$ are the corresponding coauthors. a_n can include multiple affiliations in a certain paper; however, matching the affiliation and the author in the current dataset format is impossible. c_n can also be related to multiple coauthors. Let us define *coauList* as the list of the coauthors, *affList* as the list of the affiliations, and *trupaList* as the list of the papers that have the same author p_1. This algorithm only applies to the papers written by the same author p_1 and excludes the papers that might be clustered because they are written by another author with namesake S. These three lists are recursively updated by scanning from p_2 to p_n (inner loop), and further upgrading may be done depending whether the *trupaList* between the current and the previous loop (outer loop) still grows. In the algorithm, we also set the iteration time parameter to decide whether the outer loop ends or not. The computational flow of this algorithm is shown in Fig. 3.

This algorithm has the following advanced features that enable it to detect homonymous names. First, single-author papers can be classified correctly using their affiliation information. Second, papers that have completely different coauthors can be clustered according to the same affiliation information. Third, papers whose authors have changed affiliations can be classified correctly based on the coauthorship information. This process is mainly based on our observation that authors who have new affiliation usually maintain contact with their past coauthors and collaborate continually for a period of time. Their collaborations might have started in the past affiliation and completed in the author's new affiliation, which usually happens when a Ph.D. student accepts a faculty position in another institution and still collaborates with his/her former supervisor or colleagues in the university after he/she finished doctoral studies. In the algorithm, the similarity of the authors and that of the affiliations are calculated by string matching. If the two strings of the author names are equal, they are identified as belonging to the same author name. However, the two affiliations are considered the same if the two strings of these two affiliations are over q percent by comparing one by one the upper case characters of their names. In this paper, we set threshold q as 50 %, which has been checked to be enough in the string comparison for the verification of two affiliations that are the same in our dataset. The affiliation and coauthorship information can recursively reinforce each other to identify the papers written by the same author, as shown in Fig. 4.

Using "Christopher C. Yang" as an example to illustrate how this algorithm works, the red arrows in Fig. 5 indicate the information used to disambiguate the names. We found that Yang, CC moved from Chinese University Hong Kong to Drexel University, although he still collaborates with Liu, N in Chinese University Hong Kong. Although in 2009, Yang, CC was already affiliated with Drexel University, we could still use the affiliation information "Chinese Univ Hong Kong" of his coauthors to identify his other written articles. Accordingly, the authorship of the article in 2002 can be disambiguated, as well as that of the article in 2000, although these two papers have completely different coauthors.

Revisiting the "Leydesdorff, L" example as well, the red and blue arrows in Fig. 6 indicate the application of the affiliation and coauthorship information, respectively. In 1990, Leydesdorff published an article using "Univ Amsterdam" as his affiliation; however, in 1994 and 1997, he published another articles using a different affiliation—"Dept Sci & Technol Dynam." Using only the affiliation information cannot identify the authorships of these papers. The two articles in 1990 and 1994 are single-author papers and are also impossible to identify using only the coauthorship information. Nevertheless,

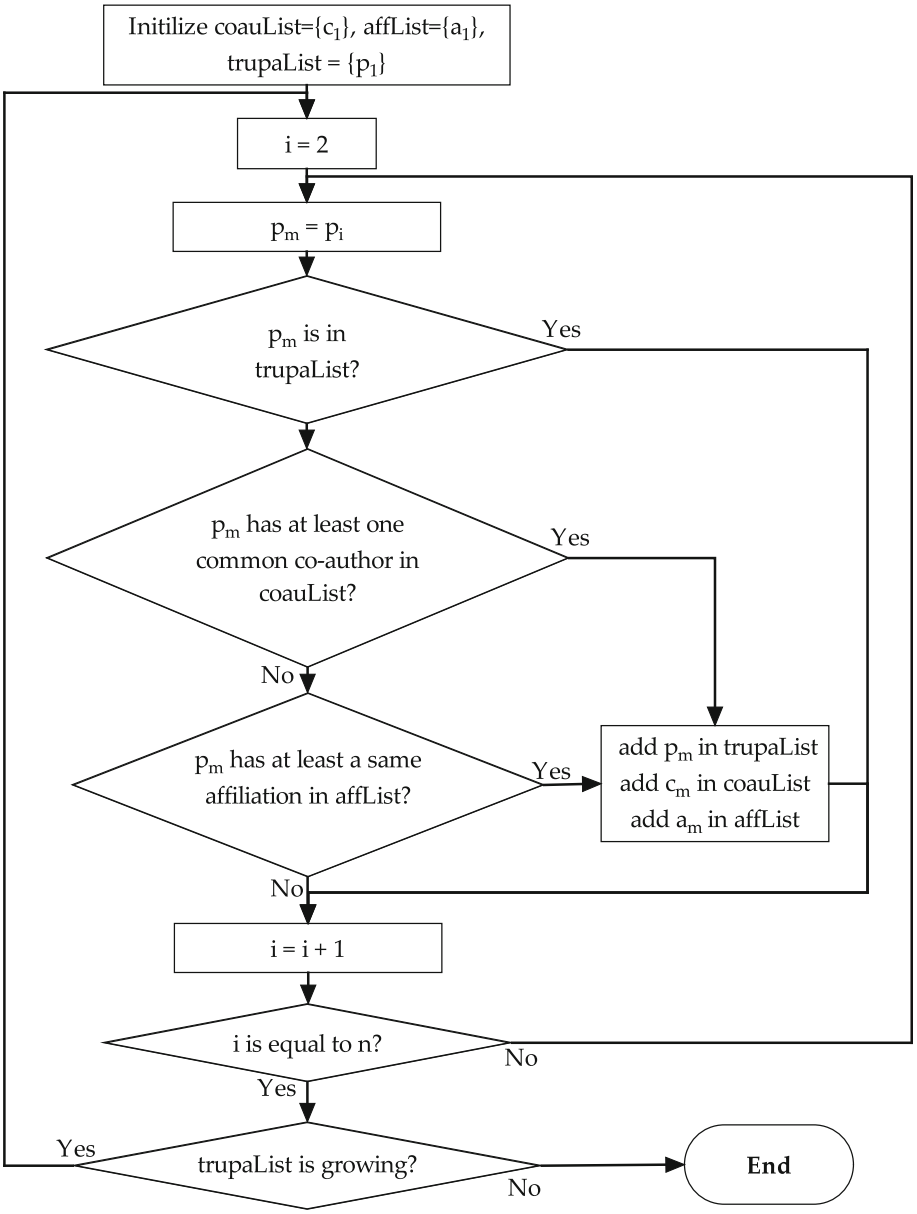

Fig. 3 Name disambiguation algorithm based on the affiliation and coauthorship information

using the proposed method as shown in Fig. 6, the article in 1996 can be identified using the affiliation "Univ Amsterdam," the article in 1997 can be identified using the coauthor "Vandenbesselar, P," and the article in 1994 can be identified using the affiliation "Dept Sci & Technol Dynam." Thus, we have shown that the affiliation and coauthorship information can reinforce each other to disambiguate the author names, and using only either one could lead to a lower recall in the name disambiguation.

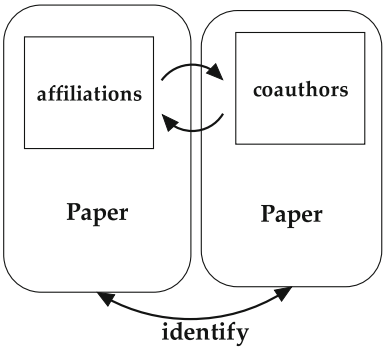

Fig. 4 Recursive reinforcement between the affiliation and coauthorship information

2009|YANG, CC;LIU, N|*Drexel Univ*|*Drexel Univ*|*Chinese Univ Hong Kong*|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY|000263935100006

2002|YANG, CC;CHUNG, A|*Chinese Univ Hong Kong*|*Chinese Univ Hong Kong*|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY|000173320900011

2000|YANG, CC;LUK, JWK;YUNG, SK;YEN, J|*Chinese Univ Hong Kong*|*Chinese Univ Hong Kong*|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE|000085224800004

Fig. 5 Illustration using the affiliation and coauthorship information for the name disambiguation using as example the case of "Christopher C. Yang"

1990|LEYDESDORFF, L|*UNIV AMSTERDAM*||SCIENTOMETRICS|A1990CY00700008

1996|*VANDENBESSELAAR, P*; LEYDESDORFF, L|*UNIV AMSTERDAM*|UNIV AMSTERDAM|JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE|A1996UM12100003

1997|LEYDESDORFF, L;*VANDENBESSELAAR, P*|*DEPT SCI & TECHNOL DYNAM*|*DEPT SOCIAL SCI INFORMAT*|SCIENTOMETRICS|A1997WP65300011

1994|LEYDESDORFF, L|*DEPT SCI & TECHNOL DYNAM*||SCIENTOMETRICS|A1994PD56500004

Fig. 6 Illustration using the affiliation and coauthorship information for the name disambiguation in the case of "Leydesdorff, Loet"

Evaluation experiments

To evaluate the proposed algorithm, we use the ISI WoS dataset. Conducting the evaluation experiments in the entire WoS dataset for all disciplines is impractical because building a ground truth to test the recall and precision of the algorithm in such a large-scale dataset is impossible (Gurney et al. 2012). In the following, we first describe the dataset for the evaluation. Second, we discuss how the iteration times are selected. Third, we compare

the situations that could possibly lead to errors when our algorithm is run. Fourth, we analyze the recall and precision of the method. Finally, we investigate the *h*-index and *g*-index distributions before and after the name disambiguation using the proposed algorithm.

Dataset

To evaluate the performance of our name disambiguation method, we download the ISI WoS dataset. By continuing to the "Advanced Search" in WoS, we use the "(SU = Information Science & Library Science)" query to search all the papers in the research area "Information Science & Library Science" of the "Social Sciences Citation Index" database. We then use the "Output Records" function at the bottom of the result webpage to download all the publication records and save them to the Tab-delimited files. Each record of the downloaded samples includes the information on the coauthors and affiliations that we will use to identify each author name. Our dataset contains 654 author names. The WoS affiliations did not individually match each author until 2009. Nevertheless, the affiliation information is still the most effective information available to disambiguate the author names. Table 1 lists the top 12 author names in terms of the number of their respective publications. They are not common names and can be identified only by manual checking based on their affiliations. However, because sometimes the same affiliation could have different names, e.g., "Univ Coll London" also uses the abbreviation UCL, developing a program to identify automatically the author names using only their affiliations is difficult. Therefore, we must integrate the coauthorship information to perform further identification. For affiliations such as "Hungarian Acad Sci" and "Lib Hungarian Acad Sci," we apply the similarity computation to decide whether they are the same.

Selection of iteration times

In our method, the algorithm processes the coauthor and affiliation information iteratively several times to solve the problems when either information is only used in the name disambiguation. We examine the process of selecting a suitable iteration time parameter. The performance is checked by the ratio of the difference between the number of publications of the author before and after the name disambiguation (*Pubnum_dis*) to that before the name disambiguation (*Pubnum_all*), as expressed in Eq. (1).

$$Performance\_ratio = (Pubnum\_all - Pubnum\_dis)/Pubnum\_all \quad (1)$$

The effect of the number of iteration times on the performance is shown in Fig. 7, where the sorted ratios for each author are shown. The second iteration improves the performance achieved in the first iteration, but the third iteration does not introduce much improvement. In particular, authors with a performance ratio of smaller than 0.2 improve much after several iteration times in the algorithm. Therefore, selecting the iteration times to be larger than three is a better option to reinforce the affiliation and coauthorship information relationship. In the evaluation experiments, we iterated the algorithm five times. Out of the 654 author names, 239 did not have name ambiguation problems, and their ratios are equal to zero. Some examples are listed in Table 2, which shows the author names and their corresponding number of publications. Table 3 shows some examples of the author names with name ambiguation problems, and the performance ratios are between 0.6 and 0.4.

Table 1 Top 12 author names in terms of the number of publications

Author name	Main affiliation	Main coauthors	Number of pub.
Tenopir, C	Univ Tennessee	King, DW; Wu, L	226
Oppenheim, C	Univ Loughborough	Ahmed, SMZ; Norris, M; Probets, S	125
Cronin, B	Indiana Univ	Meho, LI; Shaw, D; McKenzie, G	119
ROUSSEAU, R	KHBO, Hasselt Univ	Guns, R; Egghe, L; Liang, LM; Ye, FY	118
Egghe, L	Univ Hasselt, Univ Antwerp	Rousseau, R	116
Nicholas, D	Univ Coll London, UCL	Jamali, HR; Huntington, P	113
Thelwall, M	Wolverhampton Univ	Levitt, JM; Kousha, K	113
Jacso, P	Univ Hawaii, Univ Hawaii Manoa	Tiszai, J	103
McClure, CR	Florida State Univ, Syracuse Univ	Mon, L; Hernon, P; McClure, CR	89
Spink, A	Penn State Univ, Univ Pittsburgh	Zhang, Y; Jansen, BJ; Ozmutlu, S	86
Schubert, A	Hungarian Acad Sci, Lib Hungarian Acad Sci	Braun, T; Glanzel, W;	84
Leydesdorff, L	Univ Amsterdam	Egghe, L; Zhou, P; Park, HW	83

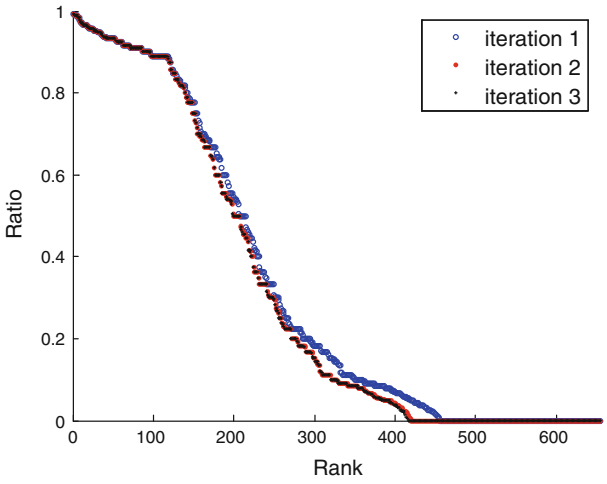

Fig. 7 Effect of the iteration times on the performance

Evaluation of the possibility of errors

The algorithm for the name disambiguation could lead to incorrect omission (false negatives or lack of recall) and incorrect assignment of the papers (lack of precision). If only the coauthorship information is used to disambiguate the author names, papers that do not share the same other author names as those of the other papers or that only have a single author will lead to low recall. Figure 8 shows that, for each author, we do pairwise comparison for all his papers and calculate the ratio of these papers that could lead to errors in all his papers, which we denote as *AURat*. The ratio of the single-author papers to all his papers are also calculated, which we denote as *SIRat*. If only the affiliation information is used to disambiguate the author names, the papers of a certain author that do not have the

Table 2 Examples of author names without name ambiguation

Glanzel, W-102	McClure, CR-89	Spink, A-86	Rousseau, R-118
Schubert, A-84	Huntington, P-68	Bates, DW-59	Cimino, JJ-50
Jarvelin, K-47	Moed, HF-46	Large, A-41	Hripcsak, G-39
Benbasat, I-39	Garfield, E-38	Dilevko, J-37	Wilson, CS-34

Note each table cell includes the author name and his corresponding number of publications

Table 3 Examples of author names with name ambiguation

Jacso, P-103	Wilkinson, D-31	McGrath, M-28	Zhang, Y-26
Stock, WG-17	Kishida, K-16	Crestani, F-16	Robbin, A-15
Seadle, M-15	Seadle, M-15	Chang, CC-14	Ho, YS-13
Kim, H-12	Liu, ZM-12	Lee, AS-9	Chen, J-9

Note each table cell includes the author name and his corresponding number of publications

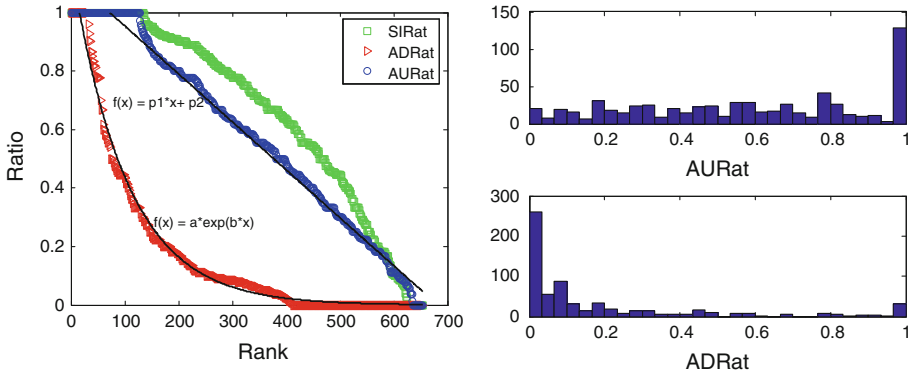

Fig. 8 Distributions of the papers that might lead to errors

same affiliation information with any other papers cannot be recalled as well. For all the papers of each author, we use pairwise comparisons to identify those that might be incorrectly omitted when only the affiliation information is considered. We denote the ratios of these papers to all the papers of each author as *ADRat*.

Figure 8 shows that the number of authors whose papers do not share the same author names as that of their other papers ($ADRat = 1$) are more than the number of authors whose papers do not share the same affiliations as that of their other papers ($ADRat = 1$). Therefore, the possibility of errors caused by using only the coauthorship information is larger than that caused by using only the affiliation information. In comparison, authors usually publish papers with similar affiliations ($ADRat = 0$), which will be useful in identifying and grouping the papers written by the same author. Furthermore, the errors that result from using only the coauthorship information in the name disambiguation are largely due to the single-author papers. Based on their distributions, ranked *ADRat* can be fitted by the exponential function $f(x) = a*\exp(b*x)$ (where the a and b parameters quantify the data), and ranked *AURat* can be described by the linear function $f(x) = c*x + d$ (where the c and d parameters quantify the data). The fits of *ADRat* and *AURat* are shown by the solid lines, where the parameters are $a = 1.168$, $b = -0.0098$,

$c = -0.001639$, and $d = 1.118$. These fits quantitatively prove the larger possibility of errors because of the coauthorship information than that of the affiliation information in the name disambiguation.

In addition, we also calculate the performance of the name disambiguation that uses the coauthorship or affiliation information only (denoted as "au" and "ad," respectively, in Fig. 9) and that where both are used (denoted by "ad & au"). The performance is calculated using Eq. (1). The method that uses only the affiliation information shows a better performance than that where only the coauthorship information is used. This performance is similar to that when both the affiliation and coauthorship information are used by a simple parallel shift from point "C" to point "B," as shown in Fig. 9. However, using only the coauthorship information omits many papers; thus, on average, the ratio is much higher, especially for authors that are ranked below 450, as indicated by point "A." After the top 450 ranks of point "A," the ratios drop linearly quickly, which indicates that the performance when only the coauthorship information is used improves and increases the recall ratio in the name disambiguation.

Overall, the possibility of errors that lead to low recall is, on average, much higher in our dataset when only the coauthorship or affiliation information is used. We need to integrate both to improve the performance of the name disambiguation.

Evaluation of recall and precision

To develop an algorithm for the name disambiguation, the main evaluation must find a baseline dataset (ground truth). Automatically establishing that the name disambiguation is correct or not is difficult, and a baseline dataset (ground truth) is necessary where the authors have already been correctly assigned. Usually, the results are checked manually to decide the veracity of the classification (Tang and Walsh 2010); however, practical applications in a larger scale database such as the WoS dataset are still not enough (Smalheiser and Torvik 2009). In this paper, we use the *random sampling* method to select the baseline dataset, check it manually, and then compare the manual results with the

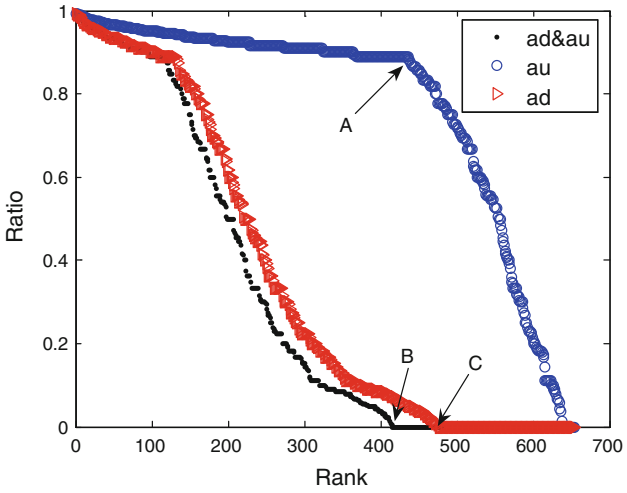

Fig. 9 Comparison of the different disambiguation methods

results obtained by the algorithm. Based on the evaluation by Onodera et al. (2011) and Gurney et al. (2012), the recall and precision ratios are defined in Eqs. (2) and (3), respectively.

$$\text{Recall ratio} = m/(m + n - p) \tag{2}$$

$$\text{Precision ratio} = m/(m + o - p) \tag{3}$$

Here, m, n, o, and p are the number of papers that the manual/algorithm identified as true/true, true/false, false/true, and false/false, respectively. We randomly sample and check 100 authors to build a baseline dataset. After the calculations, on average, the recall ratio is 92 %, and the precision ratio is 87 %.

Distribution of the *h*-index and *g*-index before and after the name disambiguation

The *h*-index and *g*-index are famous effect indicators in evaluating the scientists based on their publications, which are influenced by the author name ambiguation. The *h*-index is defined as the number of top h papers that received at least h citations and focuses on counting the number of highly cited papers whose citations do not affect the *h*-index as long as the papers are entered into the *h*-index core (Hirsch 2005). To measure the global citation performance of the author's papers, the *g*-index is proposed, defined as the largest number in which the top g papers received a total of at least g^2 citations (Egghe 2006). In this paper, the *h*-index and *g*-index distributions of all authors are investigated before and after the name disambiguation. Each author name could be related to more than one author, and thus, papers listed by an author name could be written by more than one author. Here, each calculation is only good for the papers grouped based on the first publication of the author name in his publication list.

The distributions of the *h*-index and *g*-index before and after the name disambiguation are shown in Fig. 10. In general, the *h*-index ≤ 10 and *g*-index ≤ 20 distributions cause a large change before and after the name disambiguation, and the decrease in each bar doubles the bar of the zero *h*-index or *g*-index in the histogram. The *h*-index ≥ 10 and *g*-index ≥ 20 distributions do not change significantly. However, the difference in the *h*-index or *g*-index before and after the name disambiguation could not be considered as small. Figure 11 shows that the difference in the *h*-index and *g*-index for each author before and after the name disambiguation is somewhat large, and the difference in the *g*-index is, on average, larger than that of the *h*-index. Although the *h*-index and *g*-index can exclude papers with low citations when these papers are mixed because of the name ambiguation, we must still use the name disambiguation to improve the accuracy of the calculation.

Discussions and conclusion

In reality, name disambiguation is impossible to automate. In particular, papers that do not share the same author and affiliation information with that of some other papers in the publication list of an author name cannot be identified using the proposed name disambiguation. For example, Table 4 shows the impossibility of disambiguating that a paper published in 2005 is written by the same author who published two papers in 2009.

The author name disambiguation remains an open problem and is mainly composed of three challenges (Smalheiser and Torvik 2009). First, a single individual may use different

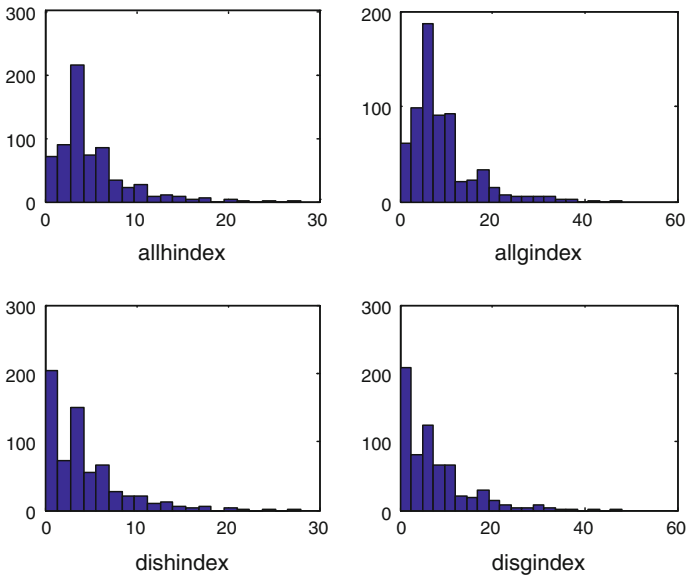

Fig. 10 Distributions of the *h*-index and *g*-index before (*allhindex* and *allgindex*) and after (*dishindex* and *disgindex*) the name disambiguation

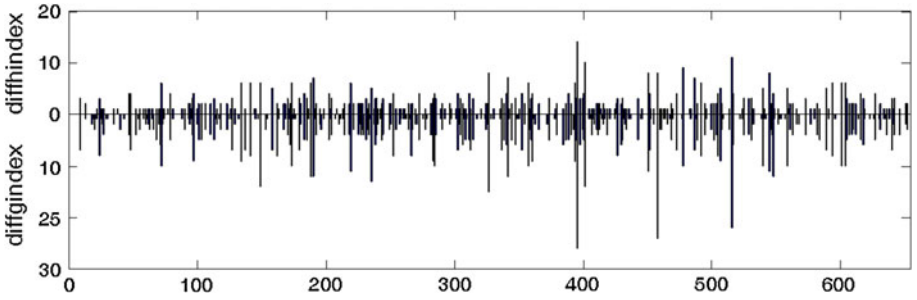

Fig. 11 Difference in the *h*-index (*diffhindex*) and *g*-index (*diffgindex*) for each author before and after the name disambiguation

names for his/her publication (e.g., the name is changed after marriage). Using only the bibliometrical data to identify the relationships between different names and a certain person is definitely not possible (Radicchi et al. 2009). Second, much information can be missing (e.g., the first and middle names are usually represented by the first letters only, resulting in insufficient information compared with the use of the full name). However, 92 % of the cases in the Physical Review publication where first letters are used in the first and middle names correspond to a single author with the same full name (Radicchi et al. 2009). Although the current paper uses a different dataset (WoS), their result can, to some extent, still be applied to our WoS dataset. Third, the same initial and last names may be used by many different individuals (e.g., some common names may be related to thousands of individuals), which generates the problem of author homonym in the name ambiguation (Kang et al. 2009). A thorough solution of this problem is still not possible because data

Table 4 Example of invalid usage of the proposed name disambiguation

2009—Bawden, D; Robinson, L—City Univ London—City Univ London—Journal of Information Science—000264022100005
2009—Robinson, L—City Univ London—City Univ London—Journal of Documentation—000269387000004
2005—Robinson, L; Hilger-Ellis, J; Osborne, L; Rowlands, J; Smith, JM; Weist, A; Whetherly, J; Phillips, R—SW London Strateg Hlth Author—SW London Strateg Hlth Author—Health Information and Libraries Journal—000235171000005

such as affiliation, e-mail, coauthor, references, and personal webpages are needed for verification, which are difficult to access and integrate in practice (Tang and Walsh 2010).

Furthermore, author disambiguation should be addressed more earnestly in any analysis at the individual level (Tang and Walsh 2010). For example, when studying collaboration networks, especially when looking for the collaborations of a certain author, author name disambiguation is necessary (Badar et al. 2012; Guns et al. 2011; Zhao and Strotmann 2011). In the influential work of the coauthorship networks by Newman (2001, 2004), the methods that minimize the bias were employed instead of author disambiguation. Even some recently published works did not apply name disambiguation (Chung and Park 2012). For some other purposes, such as the investigation of the distributions, author disambiguation is usually bypassed because the work is dedicated to study an entire problem, and the errors in the name ambiguation can be considered as randomly distributed, assuming that no name bias exists in Science (Tang and Walsh 2010). For statistical physicists who study bibliometrics (Science of Science), this principle is normally applied to avoid the problem of name ambiguation or to minimize the bias of the author identification. Name disambiguation is usually ignored, especially in statistical distribution studies (Petersen et al. 2010). One main reason is that conducting name disambiguation in a large-scale dataset is very difficult (Huang et al. 2006), and any undeveloped methods may lead to new noise in the ensuing statistical analysis (Petersen et al. 2011). For example, Petersen et al. (2010, 2011) used the author ID in the ISI WoS, which consists of the last name, first name, and middle initial. Further, Laherrère and Sornette (1998) and Radicchi et al. (2009) generated the author ID using the last name, first name, and middle initial in the Physics Review Archives. Another reason is that the purpose of these papers is to investigate the distributions of the indicators at the whole level rather than at the individual level.

In this paper, we verified the insufficiency in disambiguating the author name by considering the coauthorship or affiliation information separately. Therefore, a new name disambiguation method has been proposed by integrating this information, especially suitable in the events of scientific collaboration and mobility. The algorithm enables the coauthorship and affiliation information to recursively reinforce each other to improve the performance. Based on the results before and after the name disambiguation, the distributions of the h-index and g-index are also investigated. The results of the name disambiguation show that using only the affiliation information results in better performance than that using only the coauthorship information. However, our proposed method, which integrates both, performs better in the name disambiguation. Nevertheless, when the analysis is at the individual level rather than at the entire level, author disambiguation must be conducted seriously (Tang and Walsh 2010). Name disambiguation cannot be completely realized. Therefore, in bibliometrics, some scholars selected uncommon author names to minimize the errors caused by the third challenge of name disambiguation (Iglesias and Pecharromán 2007; Smalheiser and Torvik 2009). Using the commonness of the author names as an exclusion criterion can control the bias of name ambiguation

(Hirsch 2005; Iglesias and Pecharromán 2007). Using an uncommon author name in publications to avoid name ambiguation is also advisable, e.g., some Chinese scholars usually add an English middle or first name in the authorships of their English publications. In addition, ResearchID.com invites researchers to register for a unique researcher ID number and upload their publications. This method will be a great step in solving the problem of name ambiguation if the application for ResearchID is spread out to attract more researchers. In the future, we can also use the data from ResearchID (or any other services that allow authors to maintain their own publication lists, like Academia.edu and ResearchGate) to build a "ground truth" for disambiguation analysis, instead of manually cleaning the authorship data.

Acknowledgments This work was supported in part by the ISTIC-THOMSON Joint Scientometrics Lab Fund (Grant No. IT2012004) and in part by the China National Natural Science Fund (Grant No. 71101059).

References

Badar, K., Hite, J., & Badir, Y. (2012). Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in Pakistan. Scientometrics, 1–21, doi:10.1007/s11192-012-0764-z.

Chung, C., & Park, H. (2012). Web visibility of scholars in media and communication journals. Scientometrics, 1–9, doi:10.1007/s11192-012-0707-8.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69*(1), 131–152.

Guns, R., Liu, Y., & Mahbuba, D. (2011). Q-measures and betweenness centrality in a collaboration network: A case study of the field of informetrics. *Scientometrics, 87*(1), 133–147.

Gurney, T., Horlings, E., et al. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics, 91*(2), 435–449.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569.

Huang, J., Ertekin, S., & Giles, C. (2006). Efficient name disambiguation for large-scale databases. *Knowledge Discovery in Databases, PKDD, 2006*(4213), 536–544.

Iglesias, J., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.

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

Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B, 2*(4), 525–539.

Newman, M. E. J. (2001). Scientific collaboration networks.I. Network construction and fundamental results. *Physical Review E, 64*(1), 16131.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5200–5205.

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), 677–690.

Petersen, A. M., Jung, W., Yang, J., & Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences, 108*(1), 18–23.

Petersen, A. M., Wang, F., & Stanley, H. E. (2010). Methods for measuring the citations and productivity of scientists across time and discipline. *Physical Review E, 81*(3), 36114.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 56103.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1–43.

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

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

Wooding, S., Wilcox-Jay, K., Lewison, G., & Grant, J. (2006). Co-author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. *Scientometrics, 66*(1), 11–21.
Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis: A comprehensive comparison in the highly collaborative stem cell research field. *Journal of the American Society for Information Science and Technology, 62*(4), 654–676.