Expert Systems with Applications 39 (2012) 10521–10532

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Disambiguating authors in citations on the web and authorship correlations

Hsin-Tsung Peng [a,b], Cheng-Yu Lu [a,*], William Hsu [a,c], Jan-Ming Ho [a,b]

[a] Institute of Information Science, Academia Sinica, Taiwan
[b] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[c] Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan

ARTICLE INFO

Keywords:
Author name disambiguation
Citation analysis
Web correlation
Authorship correlation

ABSTRACT

Members of the academic community have increasingly turned to digital libraries to search for the latest work of their peers. On account of their role in the academic community, it is very important that these digital libraries collect citations in a consistent, accurate, and up-to-date manner, yet they do not correctly compile citations for myriads of authors for various reasons including authors with the same name, a problem known as the "name ambiguity problem." This problem occurs when multiple authors share the same name and particularly when names are simplified as in cases where names merely contain the first initial and the last name. This paper proposes a reliable and accurate pair-wise similarities approach to disambiguate names using supervised classification on Web correlations and authorship correlations. This approach makes use of Web correlations among citations assuming citations that co-refer on publication lists on the Web should to refer to the same author. This approach also makes use of authorship correlations assuming citations with the same rare author name refer to the same author, and furthermore, citations with the same full names of authors or e-mail addresses likely refer to the same author. These two types of correlations are measured in our approach using pair-wise similarity metrics. In addition, a binary classifier, as part of supervised classification, is applied to label matching pairs of citations using pair-wise similarity metrics, and these labels are then used to group citations into different clusters such that each cluster represents an individual author. Results show our approach greatly improves upon the name disambiguation accuracy and performance of other proposed approaches, especially in some name clusters with high degree of ambiguity.

1. Introduction

MEMBERS of the academic community have increasingly turned to digital libraries such as the Digital Bibliography and Library Project (DBLP) and Citeseer to search for the latest work of their peers or famous researchers. These digital libraries are huge collections of publications and provide bibliographical search services. On account of their role in the academic community, it is very important that these digital libraries collect citations in a consistent, accurate, and up-to-date manner, yet they do not correctly compile citations for myriads of authors for various reasons including incomplete citation information or authors with the same name (Han, Zha, & Giles, 2005).

The latter problem, known as the "name ambiguity problem," has become a major challenge in recent years because these libraries incorrectly identify various authors, incorrectly attribute academic credit to various authors, and poorly integrate data from multiple sources. This problem occurs when multiple authors share the same name and particularly when names are simplified as in cases where names merely contain the first initial and the last name. For example, Han et al. found that the publication list of "Yu Chen" on the DBLP website contained citations authored by three distinct authors with the same name (Han, Zha, & Giles, 2005). Lee, Kang, Mitra, Giles, and On (2007) also found that the publication list of "Wei Wang" contained citations authored by four distinct authors. Solving this problem consists of grouping citations in a given set of those authors sharing the same name into several clusters such that each cluster represents a distinct author. While the trouble that the name ambiguity problem poses for the academic community may seem alarming and the solution to this problem may seem easy, solving this problem is easier said than done.

Previous work on disambiguating authors, e.g. (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2005; Lee, On, Kang, & Park, 2005), was unsuccessful and focused on using the fields within citations, including the author, title, and venue. These fields of different citations that refer to the same author were thought to be strongly related, and thus, relationships could be measured using some string-based or token-based methods (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2005; Han, Xu, Zha, & Giles, 2005; Han, Zha, & Giles, 2005; Lee, On, Kang, & Park, 2005). For example, if two citations belonged to the same

[*] Corresponding author. Tel.: +886 2 27883799; fax: +886 2 27814814.
E-mail address: cylu@iis.sinica.edu.tw (C.-Y. Lu).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.121

Table 1
Citations of "John R. Smith" in DBLP website.

	Author citations
1.	**John R. Smith**, "Guest Editor's Introduction: What's New with MPEG?" *IEEE Multimedia*, vol. 12, no. 4, pp. 16–17, 2005
2.	Yi Wu, Ching-Yung Lin, Edward Y. Chang and **John R. Smith**, "Multimodal information fusion for video concept detection," *ICIP*, pp. 2391–2394, 2004

author, they likely share some common co-authors or title words. However, the information within citations might not be sufficient for correct name disambiguation. Table 1 illustrates two citations that belong to the same author "John R. Smith" but do not share any common features in the title, venue, or author fields. In this example, the relationships between citations cannot be measured due to the insufficient information within the citations. Making use of the relationships between citations requires more additional information from other resources, such as Web references, to improve the accuracy of name disambiguation (Lu, Nie, Cheng, Gao, & Wen, 2007; Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008).

Our paper proposes an effective solution to the name ambiguity problem that uses two kinds of correlations among citations: *Web Correlations and Authorship Correlations*. Web correlations depend on authors' publications being listed on their publication lists or on their faculty's publication lists on the Web. If two citations co-refer to the same author on these publication lists, they likely refer to the same author. Since publication lists contain a lot of citations, Web correlations among citations is based on co-referring citations on these Web publication lists. Authorship correlation relies on two author identification strategies. First, authorship correlations give different priority to authors' names depending on the popularity of their names. Some authors' names are rarer than others. If two citations have the same co-authors and the author's name on these citations is rare, the name likely does not belong to two different people but to the same person, so these citations should be treated as having the same author. Second, authorship correlations treat citations with the same full name, e-mail, or other "personal information" as having the same author. To test the strength of our authorship correlations, the publication pages and articles for each citation were collected from digital libraries, and then the strength of the authorship correlation was measured based on whether the citations have the same author information or not. After calculating pair-wise similarities using Web and authorship correlations, a binary classifier was applied to label matching pairs, and then these labels were used to group the citations into several clusters such that each cluster represented a distinct author. The results demonstrated our approach has a significantly higher average disambiguation accuracy than those of other approaches, especially in some name clusters with a high degree of ambiguity. The results also demonstrated that using the personal information of authors significantly improves disambiguation of authors.

The remainder of this paper is organized as follows. Section 2 reviews related works, and Section 3 describes our approach for pair-wise disambiguation. In Section 4 and 5, we discuss the experiment results. Finally, we draw conclusions in Section 6.

2. Related work

Name disambiguation is a general problem of record linkage, or linking records together such that they refer to the same entities (Treeratpituk & Giles, 2009). Much research has focused on this problem using different types of data including geographic names (Smith & Crane, 2002), biomedical terms (Al-Mubaid & Chen, 2006), and personal names (Vu, Takasu, & Adachi, 2008). This paper focuses on the problem of disambiguating authors in citations who have the same names in digital libraries, especially cases in which the authors' names are simplified. In general, the name disambiguation problem in citations is attacked by grouping citations using pair-wise similarities (Kang et al., 2009).

The pair-wise similarities in previous work, e.g., (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2005; Han, Zha, & Giles, 2005; Lee, On, Kang, & Park, 2005), were commonly made using the author, title, and venue fields of citations to identify whether two citations refer to the same author or not and made three assumptions about authors. First, they assumed that authors may publish their academic articles with the same groups of co-authors. Second, authors usually focus on the same or related research areas. Third, they usually publish their articles in the similar journals or conferences. Han, Zha, and Giles (2005), Han, Giles, Zha, Li, and Tsioutsiouliklis (2005) and Han, Xu, Zha, and Giles (2005) clustered authors, title and venue words using the same or similar concepts and improved disambiguation accuracy. Lee, On, Kang, and Park (2005) defined the name disambiguation problem as two practical problems, mixed citation and split citation, and solved these problems using author, title, and venue fields to associate the citations. Kang et al. Kang et al. (2009) explored author name disambiguation using the listed co-authors in known citations and colleagues of authors in the citations from the Web.

However, pair-wise similarities using merely the author, title, and venue fields of citations have not been as successful as those using other fields and references. A lot of research has been directed at name disambiguation, and the various approaches to name disambiguation proposed have made use of certain relationships among citations to calculate pair-wise similarities including Web references (Kanani and McCallum, 2007; Kanani, McCallum, & Pal, 2007; Lu, Nie, Cheng, Gao, & Wen, 2007; Pereira et al., 2009; Tan, Kan, & Lee, 2006; Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008), topic relationships (Song, Huang, Councill, Li, & Giles, 2007; Yang, Peng, Jiang, Lee, & Ho, 2008), author e-mail addresses and affiliations (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Huang, Ertekin, & Giles, 2007; Treeratpituk and Giles, 2009), and self-citations (McRae-Spencer and Shadbolt, 2006). Name disambiguation using Web references (Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008) searches publication lists on the Web that contain both citations of a given pair and assumes pairs of citations are authored by the same individual. Some projects (Kanani and McCallum, 2007; Kanani, McCallum, & Pal, 2007; Lu, Nie, Cheng, Gao, & Wen, 2007; Pereira et al., 2009; Tan, Kan, & Lee, 2006) have gathered publication lists available on the Web and ranked them, with personal and group publication lists having higher values than those in the digital libraries (Lu, Nie, Cheng, Gao, & Wen, 2007; Pereira et al., 2009; Tan, Kan, & Lee, 2006). Name disambiguation using topical relationships assign topics to citations and assume pairs of citations having the same or similar topic are authored by the same individual since authors often focus on specific research areas. Song, Huang, Councill, Li, and Giles (2007) used a probability distribution of topics for author name disambiguation, and this distribution followed two models: Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). In our prior work (Yang, Peng, Jiang, Lee, & Ho, 2008), we built a topic association network that identifies whether a given pair of citations has similar topics. Name disambiguation using email addresses and institute affiliations assumes pairs of citations with the same email address or affiliations are authored by the same individual. This method is highly effective because email addresses are effective personal identifiers, and affiliations typically provide the places where the authors are located. Culotta, Kanani, Hall, Wick, and McCallum (2007) looked for similarities among email addresses and institute affiliations. Name disambiguation using self-citations assumes pairs of citations with ambiguous author names are authored by the same individual if they appear in the bibliographies of authors who have the same ambiguous names because most authors tend to cite their previous work. McRae-Spencer and Shadbolt (2006) tied together citations whose authors should be the same on a citation network using self-citations.

Name disambiguation approaches using pair-wise similarities can be computed by unsupervised clustering or supervised classification. Unsupervised clustering builds a similarity function without training data and uses this function to cluster citations with ambiguous names into groups of distinct authors. Han, Zha, and Giles (2005) used a k-way clustering algorithm that was variation of the traditional k-means clustering algorithm to group citations into different clusters. Lee, Kang, Mitra, Giles, and On (2007) expressed a similarity function as a linear combination of pair-wise features with different weight. Pereira et al., 2009 applied a hierarchical clustering method, which disambiguates authors in a bottom-up fashion. In contrast to unsupervised clustering, supervised classification builds a binary classifier from training data, and this binary classifier is then used to predict relations among citations with ambiguous names. These predictions are then used to group citations into clusters. Lee, On, Kang, and Park (2005) used two binary classifiers, the hybrid Naive Bayes and Support Vector Machine (SVM), to disambiguate authors in the DBLP. Our previous work, Yang, Jiang, Lee, and Ho (2006), Yang, Peng, Jiang, Lee, and Ho (2008) and Huang, Ertekin, and Giles (2007) used the SVM binary classifier to disambiguate names. Treeratpituk and Giles (2009) conducted name disambiguation on the Medline database using the random forest model. Unsupervised clustering requires building similarity functions to integrate pair-wise similarities, whereas supervised classification requires striking a balance between quality and quantity of training and testing data.

This paper proposes a reliable and accurate pair-wise similarities approach to name disambiguation using supervised classification on Web correlations and authorship correlations.

3. Proposed approach

Our approach improves on the performance of author name disambiguation of previous approaches in two ways. First, this approach makes use of Web correlations among citations on the assumption citations that co-refer on publication lists on the Web should to refer to the same author. Second, this approach makes use of authorship correlations on the assumption that citations with the same rare author name refer to the same author. Authorship correlations also assume citations with the same full names of authors or e-mail addresses likely refer to the same author. These two types of correlations are measured in our approach using pair-wise similarity metrics. In addition, a binary classifier, as part of supervised classification, is applied to label which pairs of citations match using pair-wise similarity metrics, and these labels are used to group citations into different clusters such that each cluster represents an individual author. Our approach is as follows:

1. Use Web and authorship correlations to generate pairs of citations. The similarity scores are then calculated between any two citations in the dataset using the proposed similarity metrics.
2. Create two datasets: training data and testing data. The training data is used to create a binary classifier. Every citation is paired with another and labeled matching or non-matching. Matching pairs of citations are pairs that have the same author, and vice versa.
3. Apply the binary classifier to label matching pairs in the testing data.
4. Group the citations into the appropriate clusters using the labels. Pairs of citations labeled as matching are grouped into one cluster. Each cluster is grouped such that it should represent a distinct author.

In the following sections, we discuss our approach in detail.

3.1. Web correlation

Our approach uses the Web, a rich source of various kinds of academic publications, to disambiguate authors in citations Web correlations assume authors' citations are usually listed on their publication lists or even on faculty publication lists. Authors place their citations on the Web for different reasons. Some authors, for example, want to create a publication bibliography or a personal publication list.

Although our approach, like other approaches that have been done in previous work, uses Web correlations, our approach requires some brief explanation of how it works before the details of the approach may be discussed. First, each citation title is queried in a search engine, and a set of URLs (Uniform Resource Locators) is compiled for each citation using the retrieved URLs. Then, all these sets of URLs are complied on a list of publication list candidates. On this list of publication list candidates, two pieces of information are added next to each URL: the citation(s) to which the URL belongs and the number of citations on that list that are also found at that URL, its *DF* value. For instance, when the title of citation *A* is queried, the two retrieved URLs, *X* and *Y*, from the query are placed in its URL set, and when citation *B* is queried, the two retrieved URLs, *Y* and *Z*, are placed in its set. Since both citations are found on URL *Y*, URL *Y*'s *DF* value is 2 because it is a member of both of the URL sets of citations *A* and *B*. Next, these pieces of information are used to establish which URLs are those of publication lists. Publication lists should be those URLs whose Web pages contain a lot of citations; hence, any two URL sets containing the same URL and having high *DF* values are considered to be publication lists. Next, URLs for publication lists edited by digital libraries are removed from the list of publication list candidates since digital libraries, like the DBLP, edit some of these publication lists and sometimes, incorrectly compile citations. Finally, using this list of publication list candidates, paper titles with the same author names appearing on the same publication list candidate Web page are compiled in a database as having the same author.

3.1.1. Maximum normalized document frequency

The Maximum Normalized Document Frequency (MNDF) (Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008) metric was performed to measure the strength of the Web correlations of our approach. This metric first finds the URL sets of two given citations. The metric then selects the URL with the highest *DF* value at the intersection of those two URL sets. Given URL sets, *X* and *Y*, the MNDF is defined as follows.

$$\mathrm{MNDF}(X,Y)=\begin{cases}\frac{\mathrm{Max}_{f\in X\cap Y}(DF_f)}{\mathrm{Max}_{f\in S}(DF_f)} & \text{if } X\cap Y\neq\phi,\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where DF_f is the number of citations that contains the URL *f* in *S*. Publication list candidates *S* consists of *X* and *Y*. By applying the MNDF, the similarity score of two citations approaches 1 if they have a common URL with a high *DF* value in *S*. The *DF* value of *S* is used in the denominator for normalization and reduces the influence of noise.

3.1.2. Modified maximum normalized document frequency

Our MNDF similarity metric found that many authors have fewer citations on their publication lists than expected and two reasons may account for this phenomenon. First, authors may forget to update their publication lists, so their publication lists may not include citations of more recent work. Second, some authors may be new to the academic community and have few publications up to now. Therefore, even if two citations refer to the same author, the MNDF similarity score of these citations could be undervalued if their author has fewer citations on his or her publication list than others.

In our approach, we modified the MNDF metric to amplify the *DF* value of the numerator in (1) thereby making the MNDF metric more accurate. We modified the denominator such that it is lower than the amount it would have been using the original MNDF. The similarity score is thus higher and accounts for much of the undervaluing. However, the modified MNDF metric also amplifies noise since the denominator value was adjusted and is now lower. Given URL sets, *X* and *Y*, the modified MNDF is defined as follows.

$$\text{Modified MNDF}(X,Y) = \begin{cases} \frac{\text{Max}_{f \in X \cap Y}(DF_f)}{\text{Max}_{f \in X \cup Y}(DF_f)} & \text{if } X \cap Y \neq \phi, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where DF_f is the number of citations that contain the URL *f* in *S*.

3.2. Authorship correlation

Our approach makes use of two different authorship correlations: popularity of names and author information. In the first correlation, each instance of a popular name, a name that has more than one possible referent, is given different weight in our approach based on its popularity. Some authors have rare full and simplified names, so the citations with these names should refer to the same author. For example, the author Bjorn Kvande has a simplified name B. Kvande, which is also very rare. Only one person was found to have this simplified name on the DBLP website, so citations with this simplified name should refer to the author in question. On the other hand, the author Xiaoqing Zhu not only has a popular last name, Zhu, but has even more popular simplified name, X. Zhu. 67 people were found to have this same simplified name on the DBLP website. Citations containing the latter author's full name and especially his or her simplified name have a much lower probability of referring to the author in question and thus are less helpful in disambiguating than citations containing the former author's full or simplified names.

The second type of authorship correlation uses author information including email addresses and full names. As mentioned when discussing previous work, email addresses are effective personal identifiers since every author's personal e-mail address is different. However, email addresses are not always provided, so authorship correlations must make use of other author information that may be present. While listing names in simplified form in citations is common in the academic community, authors do typically provide their full names in the head of the article to which the citation refers. Author Anoop Gupta, for example, might be listed as A. Gupta, but a quick glance at the article reveals the full name and eliminates confusion with other authors' names, Amit Gupta and Amar Gupta. An effective approach to name disambiguation using authorship correlations must be able to make use of more than one piece of author information in citations should more effective pieces like email addresses be missing.

Our approach for measuring authorship correlations using name popularity implements a name popularity measure (NPM) metric and follows a certain procedure. First, the authors' full names are collected from several authoritative digital libraries. These libraries are authoritative because they have index pages of authors who have citations. Then, these authors' full names are clustered according to their first name initial and last name and assigned popularity values based on their simplified name. Finally, these popularity values are used to calculate the uniqueness of the authors' names via the NPM metric.

Our approach for measuring authorship correlations using author information implements an authorship correspondence measure (ACM) metric and follows a certain procedure. Take two citations by author A. Gupta for example. First, the papers' titles and the simplified name, A. Gupta, are queried in digital libraries like Google Scholar.[1] Next, the author's full name and email addresses are extracted from both the publication pages and pages of the article. Author information like the email addresses may not be found in the pages of the articles. However, like most academic articles, the first page should contain the author's full name, somewhere at the top of the publication page. Last, this author information—email address, full name, or both—is used to identify whether these two citations refer to the same author or not through the ACM metric.

3.2.1. Name popularity measure

The NPM metric assigns similarity scores to simplified author names based on their uniqueness. Any two citations sharing the same rare simplified author name receive higher similarity scores and are treated as having the same author. The NPM assigns a similarity score approaching 1 if one of their authors is found to have a rare simplified name. Given two author names sets of two citations, *X* and *Y*, the NPM metric is defined as follows.

$$\text{NPM}(X,Y) = \begin{cases} 1 - \log_{\underset{f \in U}{\text{Max}}(C_f)} \underset{f \in X \cap Y}{\text{Min}}(C_f) & \text{if } X \cap Y \neq \phi, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where C_f is popularity value of simplified name *f*, i.e., the number of authors having the same simplified name *f*, and *U* is the popularity set of the dataset.

3.2.2. Author correspondence measure

The ACM metric uses author information to identify whether two given citations refer to the same author or not. Given two citations *X* and *Y*, their full name sets *A* and *B*, and e-mail sets *C* and *D*, the ACM metric is defined as follows.

$$S_{\text{full name}}(A,B) = \begin{cases} 1 & \text{if } A \cap B \neq \phi, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

$$S_{\text{e-mail}}(C,D) = \begin{cases} \frac{1}{1+e^{-|C \cap D|}} & \text{if } C \cap D \neq \phi, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

$$\text{ACM}(X,Y) = \frac{1}{2} \cdot S_{\text{full name}}(A,B) + \frac{1}{2} \cdot S_{\text{e-mail}}(C,D), \tag{6}$$

where $|C \cap D|$ is the number of e-mails at the intersection of *C* and *D*. By applying the ACM, the similarity score of the two citations approaches 1 if they have at least one common full name and more than one common e-mail. Eq. (4) is defined rather simply as the score of two citations having at least one common full name. Eq. (5) is based on the number of common e-mail addresses in the two citations and reduces the influence of incorrect calculations when e-mail address sets contain incorrect e-mail addresses. Incorrect e-mail addresses may perhaps have been extracted from the middle part of an article. Finally, these two similarity scores are added and weighted in (6).

[1] Google Scholar http://scholar.google.com.

3.3. Grouping citations with binary classifier

These two types of correlations, Web and authorship, are measured in our approach using pair-wise similarity metrics. After pairs of citations are generated and split into training and testing data, they are sorted by way of a supervised classification method. A binary classifier, as part of supervised classification, is applied to label matching pairs of citations using pair-wise similarity metrics. If the two citations of a pair refer to the same author, the pair is labeled a match; on the other hand, if the two citations of a pair refer to different authors, the pair is labeled a non-match.

Since our approach uses pair-wise disambiguation, the data are unbalanced because the datasets used by the binary classifier have a larger number of unmatched pairs than matched pairs. For instance, a set of 10 citations has 45 pairs but only 1 pair may refer to the same author and thus, be labeled a match. As the number of citations increases, the disproportion between matching and non-matching pairs increases. In some binary classifiers, like support vector machines (SVM), this disproportion can be compensated by using different penalty parameters in the SVM formulation (Osuna, Freund, & Girosi, 1997). In other words, different weights ratio C_{-+}: C_{+-} are set for the SVM classifiers to deal with disproportion. Matched pairs are classified as positive and non-matched pairs are classified as negative.

The pairs of citations labeled as matches are grouped by constructing a graph in which a vertex represents a citation, and an edge represents matching. If the pair of citations is labeled as matching, two vertices are connected. Connected vertices on the graph are deemed clusters, and each cluster should thus represent a distinct author.

4. Experiments

We conducted a number of experiments to test whether our approach improves upon the performance of other name disambiguation approaches. The experiment design, evaluation method, experiment results, and discussion are presented in the following subsections.

4.1. Datasets and experiment settings

The dataset constructed by Han, Zha, and Giles (2005) were used in our experiments because they feature ranked ambiguous name clusters In the construction of the dataset containing various citations downloaded from the DBLP website, Han et al. first clustered the citations whose author names had the same first name initial and the same last name and then ranked the name clusters according to the number of individual authors contained. The highest ranking name clusters were mostly Romanized Asian names including "C. Chen", "J. Lee", "S. Lee", and "Y. Chen." Han et al. also created 10 ambiguous name clusters and added these 14 name clusters into the dataset, as shown in Table 2. The authors' names clustered in the dataset were simplified to merely the first name initial and last name. We evaluated the uncertainty of all of the name clusters within the dataset by calculating Shannon entropy (Shannon, 1948). Since entropy value is a measure of uncertainty, the name clusters in the dataset with high entropy values are more ambiguous than those with lower entropy values. In Table 2, we can see that the name clusters "A. Gupta", "C. Chen", "J. Lee", "J. Martin", "S. Lee", and "Y. Chen" have higher degree of ambiguity than others.

We first made Web correlations using this dataset in our experiments. In making these Web correlations, first, we queried each title of the citations in Google[2] and collected the first 1000 URLs. Next, wanting only publication lists edited by authors or faculties, we filtered out URLs of digital libraries using a simple method—keyword matching. URLs containing specific keywords such as "dblp", "docis", and "/db/", for example, were filtered out. Finally, the remaining URLs were compiled on a list of valid sources for Web correlation.

Table 2
The DBLP datasets.

	Name	*N*	*C*	Distribution of citations: (#*C*):#*N*	Entropy
1.	A. Gupta	26	577	(2–10):15, (11–20):1, (21–30):3, (31–40):2, (41–50):2, (61–70):1, (91–100):1, (101–110):1	0.413
2.	A. Kumar	14	244	(2–10):9, (11–20):1, (21–30):2, (41–50):1, (91–100):1	0.347
3.	C. Chen	61	800	(2–10):42, (11–20):9, (21–30):3, (31–40):1, (41–50):2, (51–60):2, (71–80):1, (101–110):1	0.516
4.	D. Johnson	15	368	(2–10):9, (11–20):1, (21–30):1, (31–40):3, (181–190):1	0.301
5.	J. Lee	100	1417	(2–10):64, (11–20):12, (21–30):7, (31–40):7, (41–50):4, (51–60):4, (71–80):1, (81–90):1	0.554
6.	J. Martin	16	112	(2–10):12, (11–20):3, (21–30):1	0.524
7.	J. Robinson	12	171	(2–10):6, (11–20):3, (21–30):2, (41–50):1	0.409
8.	J. Smith	30	927	(2–10):18, (11–20):3, (21–30):2, (31–40):1, (61–70):1, (91–100):1, (101–110):2, (151–160):1, (171–180):1	0.371
9.	K. Tanaka	10	280	(2–10):5, (11–20):1, (31–40):2, (61–70):1, (101–110):1	0.305
10.	M. Brown	13	153	(2–10):8, (11–20):2, (21–30):2, (41–50):1	0.436
11.	M. Jones	13	259	(2–10):7, (11–20):1, (31–40):2, (41–50):2, (51–60):1	0.385
12.	M. Miller	12	412	(2–10):7, (11–20):2, (21–30):1, (141–150):1, (191–200):1	0.228
13.	S. Lee	83	1457	(2–10):49, (11–20):11, (21–30):6, (31–40):6, (41–50):5, (51–60):1, (61–70):2, (71–80):2, (191–200):1	0.509
14.	Y. Chen	71	1264	(2–10):49, (11–20):6, (21–30):7, (41–50):2, (51–60):1, (61–70):2, (71–80):1, (91–100):1, (111–120):1, (221–230):1	0.471

N denotes the number of individuals, *C* denotes the number of citations, (#*C*) indicates the range of the number of citations, #*N* is the number of individuals whose citations are within the range (#*C*), and Entropy represents the distribution of individuals' citations.

We then made authorship correlations using the dataset. In order to calculate the NPM similarity scores, we collected all of authors' names from the author index page on the DBLP website. Then, we clustered them using the first name initial and last name. Next, we calculated the popularity values of these simplified names, which are the number of author names contained in each name cluster. For example, the simplified name "S. Ranjan" had a popularity value of 4 because its name cluster contained four author names: "Supranamaya Ranjan", "Shrish Ranjan", "Sanjiv Ranjan", and "Sohan Ranjan". In order to calculate the ACM similarity scores, we queried the title and simplified author name of each citation in Google Scholar and collected the first 100 URLs for their Web pages and articles. We extracted author information, author full names and e-mail addresses, from both the Web pages and articles in such formats as PDF and PostScript files but only articles from famous digital libraries including ACM portal, IEEE Xplore, and SpringerLink. We only extracted the full name that matched the given simplified name. Since the dataset from Han et al. containing various citations downloaded from the DBLP website had some spelling errors, we did not extract the corresponding full names from Web pages and articles. For example, after crawling articles using the simplified author name "A. Gupta," we found several authors including "Anoop Gupta", "Yong Rui", and "Alex

[2] Google http://www.google.com.

Acero," had written one of the articles so only "Anoop Gupta" was extracted from that article because it matched the simplified name "A. Gupta".

We were not able to gather the author information in some of the citations for a few reasons. First, some Web pages and articles did not have author information. Second, some Web pages and articles could not be downloaded without permission. Third, some articles were images or damaged. Table 3 shows the statistics for parsing author information and demonstrates that only 74.17% (6261/8441) of citations with author information could be used in our experiments.

To deal with the unbalanced data problem for classification, we adopted the C-SVC binary classifier with an RBF (radial basis function) kernel function, implemented by LibSVM (Chang & Lin, 2001), which is the weighted SVM for unbalanced data, and a leave-one-out training schema. When one name cluster was used for testing data, the others were used for training data. In the configuration of SVM, the cost parameter C and parameter γ were set as the default values of LibSVM, and the weight ratio C_{-+}: C_{+-} between positive (matching) and negative (non-matching) classes was scanned from 1:1 to 20:1 by using grid parameter search. Given a weight ratio, an average F-measure value was calculated after all of the name clusters had been tested. After scanning all the weight ratios, the weight ratio with the highest average F-measure value was selected for SVM.

Table 3
Statistics of parsing author information.

Name	C	Number of citations with Web pages or articles (D)	Number of citations with e-mails (E)	Number of citations with full name (F)	$E \cap F$	$E \cup F$
1. A. Gupta	577	528	309	481	294	496
2. A. Kumar	244	224	112	201	106	207
3. C. Chen	800	679	296	564	271	589
4. D. Johnson	368	341	127	306	120	313
5. J. Lee	1417	1241	550	970	485	1035
6. J. Martin	112	97	48	83	45	86
7. J. Robinson	171	149	38	127	36	129
8. J. Smith	927	808	405	700	374	731
9. K. Tanaka	280	226	82	184	74	192
10. M. Brown	153	138	81	129	77	133
11. M. Jones	259	233	125	208	118	215
12. M. Miller	412	335	163	264	138	289
13. S. Lee	1457	970	543	836	502	877
14. Y. Chen	1264	1105	565	912	508	969
Total	8441	7074	3444	5965	3148	6261

C denotes the number of citations, $E \cap F$ denotes the number of citations having both e-mails and full name, and $E \cup F$ denotes the number of citations having either of e-mails and full name.

4.2. Experiment design

We designed several experiments to test our approach. First, we applied receiver operating characteristic (ROC) curve analysis to compare the effects of the similarity metrics in Web and authorship correlations. Second, we compared the performances of our approach in various combinations of similarity metrics based on their disambiguation accuracy and F-measure values. Finally, we also compared the performance of our disambiguation approach with those of Han, Zha, and Giles (2005) and our prior work (Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008).

4.3. Performance evaluation method

We tested our approach's clustering precision, clustering recall, F-measure, and disambiguation accuracy. First, we assigned a cluster to the referent individual author that appears the most in the citations in that cluster. For example, a cluster containing 10 citations of author A and 5 citations of author B would be assigned to author A. To evaluate the clustering precision and recall of our approach, we multiplied each cluster's precision and recall by a weight factor. In this test, we looked merely at the overall precision and recall since the influence of different clusters should not be equal. Larger clusters should be more influential than smaller ones. We then summed these figures of each cluster. They are defined as follows.

$$\text{Precision}_{\text{cluster}} = \sum_{g \in G} \frac{n_g}{N} \cdot \frac{n_{ig}}{n_g}, \tag{7}$$

$$\text{Recall}_{\text{cluster}} = \sum_{g \in G} \frac{n_g}{N} \cdot \frac{n_{ig}}{n_i}, \tag{8}$$

where G is the set of citation clusters in the result; n_g is the number of citations in cluster g; N is the total number of citations in the name cluster; n_{ig} is the number of citations associated to author i in cluster g, which is assigned to i; and n_i is number of citations authored by i.

F-measure measures the overall clustering precision and recall and is defined as the harmonic mean of them. It is calculated as follows.

$$F\text{-measure} = \frac{2 \cdot \text{Precision}_{\text{cluster}} \cdot \text{Recall}_{\text{cluster}}}{\text{Precision}_{\text{cluster}} + \text{Recall}_{\text{cluster}}}. \tag{9}$$

We did not evaluate disambiguation accuracy the same way as Han et al. did in their work (Han, Zha, & Giles, 2005) because our approach uses a different clustering technique. Their approach uses unsupervised clustering and requires defining the number of clusters before sorting citations into author name clusters. For example, if "A. Gupta" is the simplified name of 26 different author names in the academic community, their disambiguation approach must create 26 clusters for each of the different author names for "A. Gupta" before it sorts citations into each author name cluster. This approach is thus not practical because it is impossible to know prior to clustering how authors share the same simplified names in the real world. Hence, in our approach, we employed a binary classifier and a simple graph grouping approach. Using the same method to evaluate disambiguation accuracy as Han et al. did would have led to a few problems, so we made few modifications to their approach. First, we only selected the cluster with maximum clustering recall as the cluster for each author. As mentioned previously, the clusters are assigned to the referent individuals that appear the most in the citations in those clusters. Some clusters could be referred to the same author. For example, our approach clusters the citations of A. Gupta into 40 clusters. The citations of A. Gupta have 26 authors, as shown in Table 2. Several clusters could be assigned to one author, and we only selected one cluster with maximum clustering recall as his or her cluster. Second, we calculated the disambiguation accuracy by dividing the sum of citations in clusters with maximum clustering recall by the total number of citations in the set of clusters. Since each author can at most refer to one cluster, other clusters belonging to him or her cannot be used to calculate disambiguation accuracy. The value of numerator was decreased, and disambiguation accuracy could be influenced. The disambiguation accuracy is defined as follows.

$$\text{Accuracy} = \frac{\sum_{i \in I} n_{ir}}{N}, \tag{10}$$

where I is the set of authors in the name cluster, r is the correct cluster of author i, and N is the total number of citations in the name cluster.

4.4. Experiment results

We conducted three different evaluations: similarity metric analysis, performance evaluation using combinations of different similarity metrics, and baseline performance evaluation. The experiment results are discussed in following subsections.

4.4.1. Similarity metric analysis

We evaluated and compared the effects of similarity metrics using Web and authorship correlations using a simple clustering method to cluster the citations of the name cluster instead of our proposed approach. Since binary classifiers were not applied in this evaluation, pairs of citations were labeled matches if their similarity scores were higher than the given threshold. These citations were grouped by constructing a graph in which a vertex represents a citation and an edge represents matching, connected citations. The lower the threshold, the more citations were connected to each other. Thus, the clustering precision is low when the clustering recall is increased. The lower the threshold, the lower the clustering precision but the higher the clustering recall, and vice versa.

We used the Modified Sigmoid Function (MSF) and Cosine Similarity Metric (CSM) metrics from our prior works (Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008) and the NPM to determine whether match citations to their correct author. Fig. 1 illustrates the average ROC curve of MSF, CSM, and NPM among 14 name clusters, where the threshold ranges from 0 to 1 and each step is 0.05. We observed that NPM metric yielded better clustering recall than MSF and CSM metrics. However, in some cases, especially when the threshold was lower than 0.4, the NPM metric had lower clustering recall than the MSF and CSM metrics because author names in the dataset might have been incorrect or misspelled and did not have popularity values. Citations with such author names were not connected to any other citations when graphed and were clustered into several single clusters. These situations influenced clustering recall.

We proposed a modified similarity metric, named Modified MNDF, to test the influence of similarity metrics on Web correlation since MNDF metric might undervalue the authors who have fewer citations. Fig. 2 shows the average ROC curves of MNDF and Modified MNDF metrics. The figure reveals that the Modified MNDF metric yielded better clustering recall and lower clustering precision than the original MNDF metric. The Modified MNDF metric clusters citations with lower *DF* because it replaces the denominator of the MNDF metric and increases the clustering recall. However, it also may amplify the influence of noise during which the clustering precision decreases. On the other hand, the MNDF metric might eliminate the influence of noise and lower the *DF* values of correlations of pairs of citations as expressed in (1) thereby decreasing the clustering recall. In the following evaluations, we will discuss whether these two similarity metrics are complementary or not.

As shown in Fig. 3, which illustrates the average ROC curves of four similarity metrics used in this paper, the ACM metric did not perform better than other similarity metrics because only 6261 citations with author information (about 74.17%) could be clustered by leveraging the ACM calculation. The remaining citations could not be connected with other citations in graph since they did not have author information. The clustering recall was thus obviously limited. We removed the citations for which no author information was found and used the remaining citations to analyze their ROC curves again in order to analyze the effect of ACM metric. The results are shown in Fig. 4. We see that the ACM metric has similar clustering precision to other similarity metrics but higher clustering recall. In addition, the ACM metric has the best *F*-measure (about 0.84) when compared to the Modified MNDF (about 0.75), MNDF (about 0.70), and NPM (about 0.59) and this *F*-measure indicates that the ACM metric has the significant effect and plays an important role in author name disambiguation.

4.4.2. Performance evaluation over combinations

As shown in Fig. 5, which represents the performances for single similarity metrics using a binary classifier, the Modified MNDF metric performs better, on average, than other similarity metrics. In the case of the name cluster “Y. Chen”, the Modified MNDF metric improved upon the *F*-measure of the others by almost 20%. Although only 74.17% citations have author information, the ACM metric still performed better than NPM metric especially in the cases of “A. Gupta”, “A. Kumar”, and “J. Martin.”, and performed, on average, as well as MNDF metric. However, all similarity metrics did not perform well when it came to the name cluster “S. Lee”. Since, as shown in Table 2, this name cluster had higher entropy value in our dataset, perhaps no single similarity metrics could provide enough information to properly train the binary classifier.

As shown in Fig. 6, which compares the performances of the combinations of two similarity metrics, we observed that all combinations, except “NPM+ACM,” performed equally well. Since MNDF and Modified MNDF yielded better effects during similarity metric analysis, we expected that these similarity metrics would the most helpful for author name disambiguation. The combination “MNDF+Modified MNDF” indeed improved upon the individual performances of “MNDF” and “Modified MNDF” shown in Fig. 5, especially on name clusters “A. Gupta”, “J. Lee”, “J. Martin”, and “S. Lee”. As mentioned in Section 4.4.1, these two similarity metrics complement each other. Because citations lacking author information had been removed prior to this evaluation, the ACM metric had a limited effect on these results.

Fig. 7 illustrates a performance comparison of the combinations of any three similarity metrics. Although these combinations performed similarly on *F*-measure, they performed differently on some name clusters, especially “A. Gupta”, “S. Lee”, and “Y. Chen”. For example, two combinations “NPM + MNDF + ACM” and “MNDF+Modified MNDF+ACM” perform better than the others on “A. Gupta” and “S. Lee”, but not in “Y. Chen”.

As shown in Fig. 8, which illustrates the performance of the combination of all four similarity metrics compared to the combination of any three similarity metrics, the quadruple combination (about 0.87) performs better than triple combinations in most name clusters. This quadruple similarity metric combination improves upon the performance of single and triple combination on “A. Gupta”, “S. Lee”, and “Y. Chen”.

4.4.3. Performance evaluation with baselines

When our results were compared with those of Han, Zha, and Giles (2005) and our prior works (Yang, Jiang, Lee, & Ho, 2006; Yang, Peng, Jiang, Lee, & Ho, 2008), as shown in Fig. 9, the average disambiguation accuracy of our approach was better than those achieved by baselines where the minimal number of citations of each author is larger than 2, which were the results of Han et al. As described in Section 4.1, some name clusters with higher entropy values including “A. Gupta”, “C. Chen”, “J. Lee”, “J. Martin”, “S. Lee”, and “Y. Chen were more ambiguous than other name clusters. We observed that the disambiguation accuracy of our approach improved upon the performance of others on name clusters “A. Gupta”, “A. Kumar”, “J. Lee”, and “S. Lee”. However, the disambiguation accuracy of our approach for “C. Chen” was worse than those achieved by baselines because the contribution of similarity scores to the disambiguation of various name clusters varies. In other words, the usefulness of similarity metric relationships may be reduced for some name clusters when the binary classifier is trained for general use.

Although, as indicated by the *F*-measures in Fig. 10, our approach improved upon the performance of others substantially,

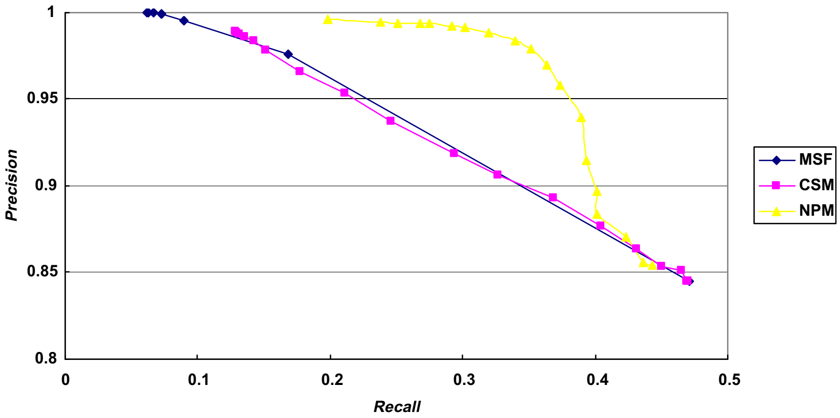

Fig. 1. The ROC curves of MSF, CSM, and NPM metrics.

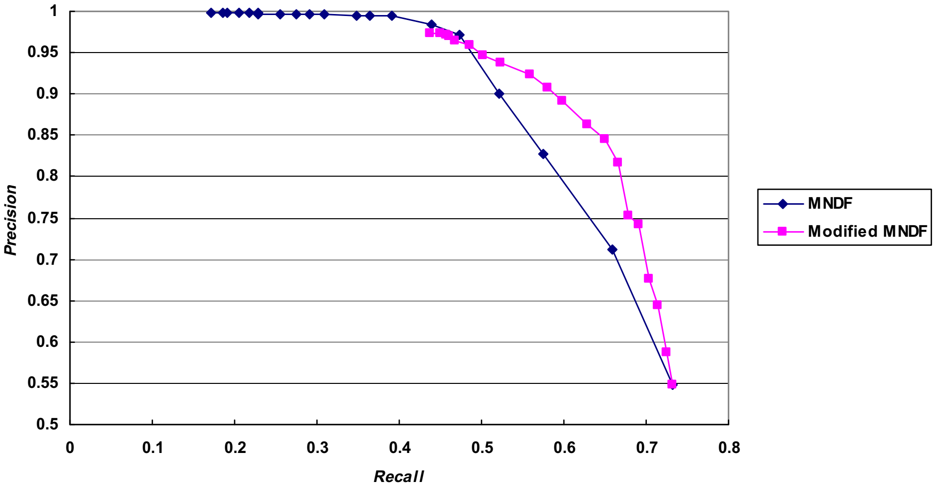

Fig. 2. The ROC curves of MNDF and Modified MNDF metrics.

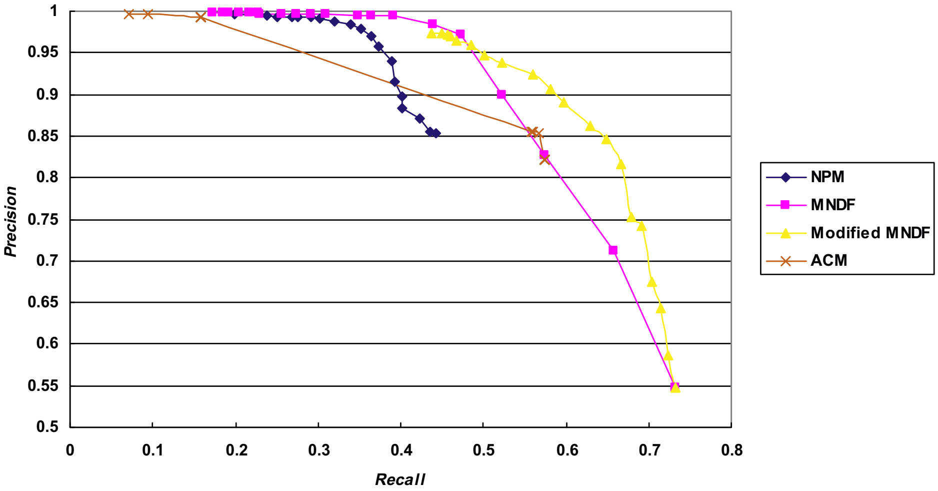

Fig. 3. The ROC curves of four similarity metrics.

especially in name clusters “A. Gupta”, “A. Kumar”, “J. Lee” and “S. Lee,” it performed worse than the others on several name clusters but only by at most 2%. In summary, leveraging both Web and authorship correlations, our approach performs well on name clusters with high entropy.

5. Discussion

In this section, we discuss some issues of our approach and illustrate them with simple examples. First, the dataset we used had some labeling errors in citations. For example, two citations “A. Gup-

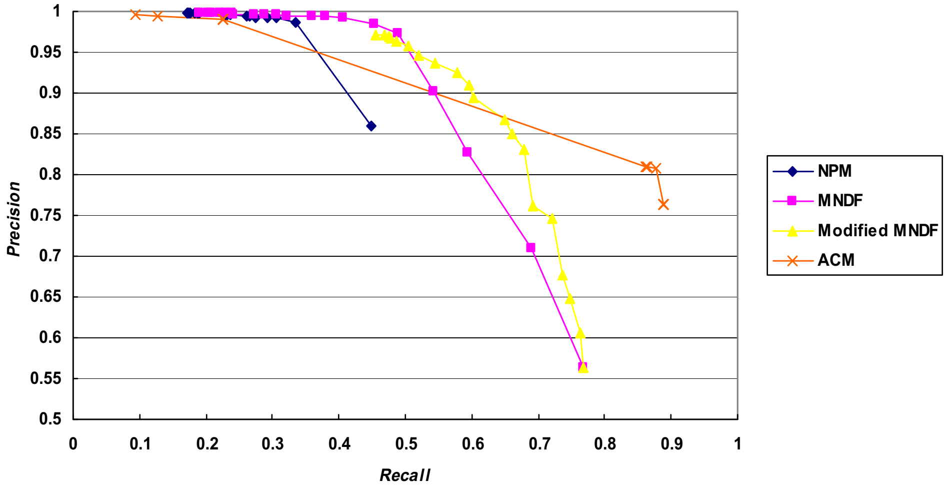

Fig. 4. The ROC curves of four similarity metrics using only the citations with author information.

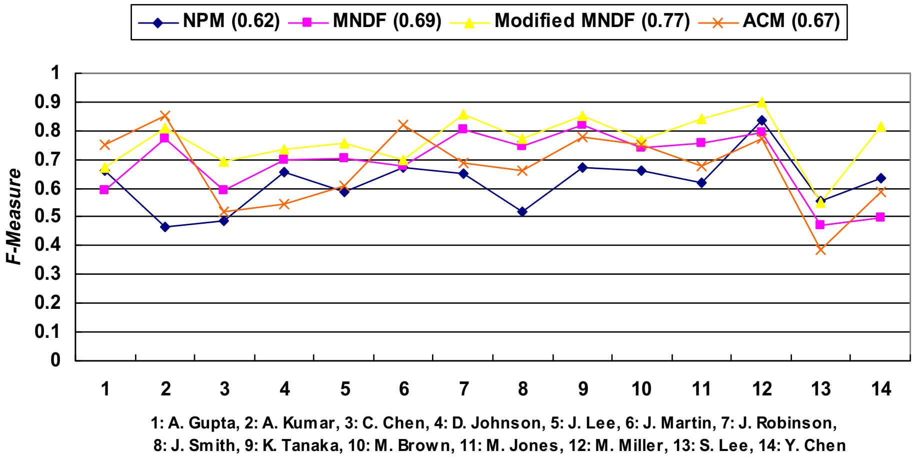

Fig. 5. *F*-measure for combinations of single similarity metrics. Note that the number in the parentheses is the average *F*-measure value among 14 name clusters.

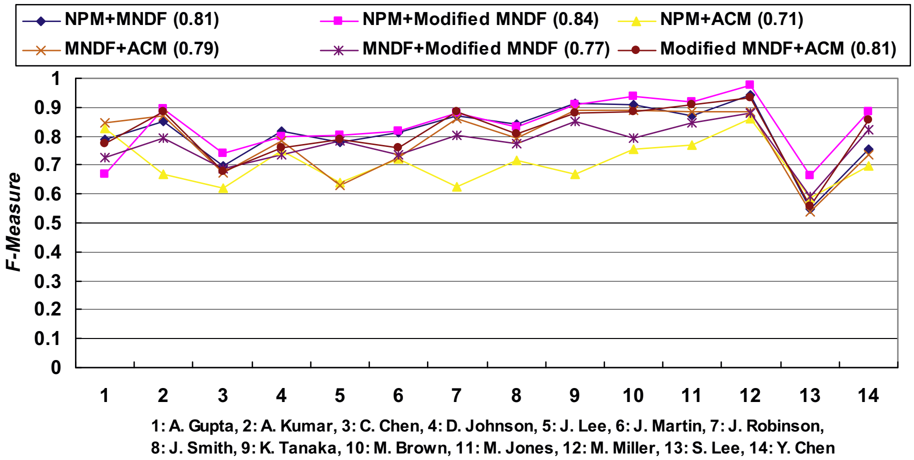

Fig. 6. *F*-measure for combinations of two similarity metrics. Note that the number in the parentheses is the average *F*-measure value among 14 name clusters.

ta, Murali Vemulapati, Ram D Sriram, Incremental Loading in the Persistent C++ Language E, JOOP Journal of Object Oriented Programming" and "A. Gupta, B.E. Prasad, M.P. Reddy, P.G. Reddy, A Methodology for Integration of Heterogeneous Databases, IEEE Trans Knowl Data Eng" were labeled as having different authors but were later found to be authored by the same author: "Amar Gupta", a co-director of productivity from the Information Technology (PROFIT) research initiative at MIT. Moreover, Pereira et al. (2009) also found that the dataset had incorrect citations and that some of them could not be found on DBLP website or other websites on the Web.

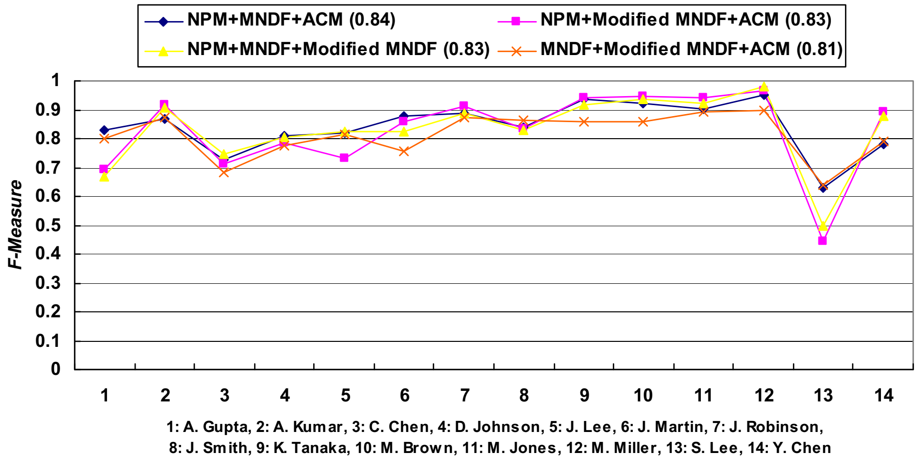

Fig. 7. *F*-measure for combinations of three similarity metrics. Note that the number in the parentheses is the average *F*-measure value among 14 name clusters.

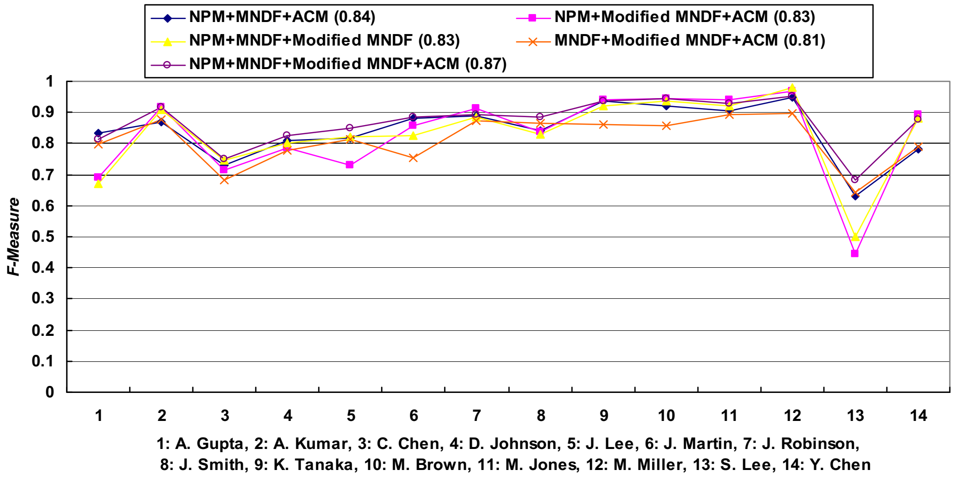

Fig. 8. *F*-measure for combinations of four similarity metrics. Note that the number in the parentheses is the average *F*-measure value among 14 name clusters.

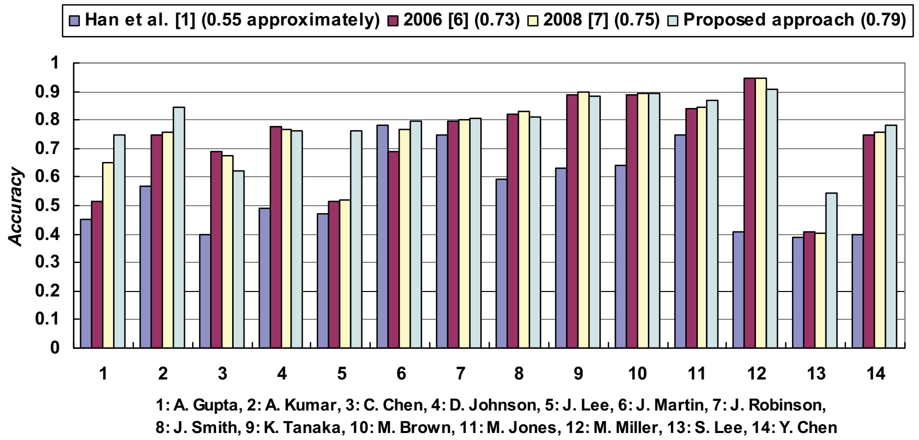

Fig. 9. Disambiguation accuracy evaluation with Han's work and our prior works. Note that the number in the parentheses is the average disambiguation accuracy value among 14 name clusters.

These problems affected the performance of our approach and led lower performance in the evaluations.

Second, we measured the relationships between citations using publication lists edited by authors or faculties on the Web since using the Web pages of digital libraries may have led to incorrect labeling on citations. However, digital libraries also provide lots of useful information. The information used for Web correlation thus had to be weighed depending on its source. Adjusting this weight system may improve our approach's performance on Web correlations.

Third, when assigning popularity values to all of the simplified names from DBLP website in order to measure the uniqueness of

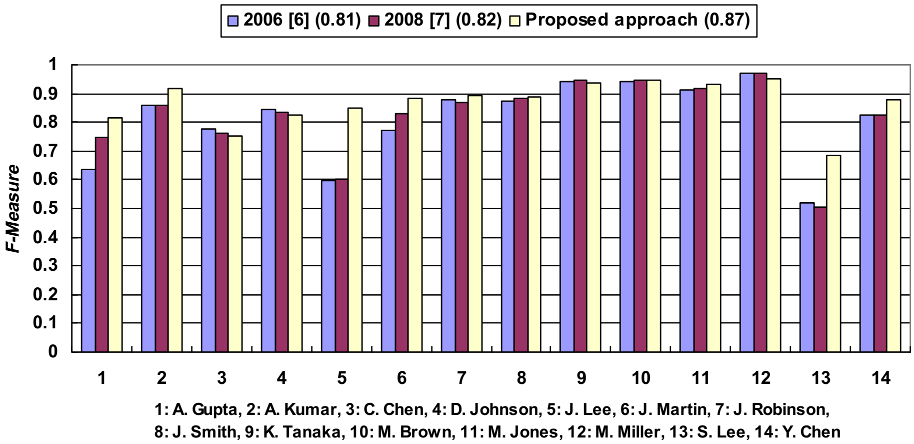

Fig. 10. *F*-measure evaluation with our prior works. Note that the number in the parentheses is the average *F*-measure value among 14 name clusters.

authors' names, we found some author names in the dataset had been misspelled in citations. For example, the citation ‘‘A. Gupta, Alaa Youssef, C. Michael Overstreet, Emilia Stoica, Hussein M. Abdel-Wahab, J. Christian Wild, Kurt Mary, The software architecture and interprocess communications of IRI: an Internet-based interactive distance learning system, WETICE Workshop Enabling Technologies Infrastructure for Collaborative Enterprises". The author name ‘‘Kurt Mary" should be spelled ‘‘Kurt Maly" according to the DBLP website. These misspellings affect the calculation of popularity values of simplified names and leads incorrect measurements of the NPM attribute.

Finally, we only crawled articles from several famous digital libraries including the ACM portal, IEEE Xplore, and SpringerLink, and we could only find author information for 74.17% of the citations in the dataset used to make authorship correlations. If our approach were to crawl more articles in such digital libraries as Citeseer and those on personal and group websites, it would be enhanced and perform even better since as described in Section 4.4.1, author information plays such an important role in author name disambiguation.

6. Conclusion

Our approach makes use of Web and authorship correlations among citations in order that their authors' names, simplified or not, can be disambiguated. Our approach first finds publication lists on the Web that were edited by authors or faculties and then treats citations with ambiguous author names listed on the same publication lists as referring to the same author. Our approach assigns popularity values to each author name and treats citations with the same rare author names as having the same author. Furthermore, our approach treats citations with the same full name or e-mail address as having the same author since authors' full names and e-mail addresses serve as effective identifiers.

Our results show that our approach clusters citations of the same author more accurately than previous approaches. Our approach greatly improves on the baselines specified in this paper, especially in some name clusters with high degree of author name ambiguity. These results can be attributed to the quality of the Web and authorship correlations and the construction of author information they provide in name disambiguation.

In future work, we plan to modify the way in which Web correlations are measured by scaling websites that were filtered out in this paper and propose a process that weighs different websites differently. Furthermore, we also plan to modify our approach such that it extracts more complete author information, such as author full names, e-mail addresses, and affiliations, for each citation in order to improve authorship correlations and thus disambiguate ambiguous authors' names in citations more accurately.

Acknowledgement

This work was supported by the National Science Council of Taiwan under Grants NSC 98-2221-E-001-010-MY3.

References

Al-Mubaid, H. & Chen, P. (2006). Biomedical term disambiguation: An approach to gene-protein name disambiguation. In *Proceedings of the international conference of information theory: New generation* (pp. 606–612).

Chang, C. & Lin, C. (2001). Libsvm: A library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm.

Culotta, A., Kanani, P., Hall, R., Wick, M. & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. In *Proceedings of AAAI-07 workshop on information integration on the Web*.

Han, H., Giles, C.L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2005). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the ACM/IEEE-CS joint conference on digital libraries* (pp. 296–305).

Han, H., Xu, W., Zha, H., & Giles, L. (2005). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In *Proceedings of the ACM symposium on applied, computing* (pp. 1065–1069).

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the ACM/IEEE-CS joint conference on digital libraries* (pp. 334–343).

Huang, J., Ertekin, S., & Giles, C. L. (2007). Efficient name disambiguation for large scale databases. In *Proceedings of the European conference on principles and practice of knowledge discovery in database* (pp. 536–544).

Kanani, P. & McCallum, A. (2007). Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes. In *Proceedings of AAAI 2007 workshop on information integration on the Web* (pp. 38–43).

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the Web. In *Proceedings of international joint conference on artificial intelligence* (pp. 429–434).

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

Lee, D., Kang, J., Mitra, P., Giles, C. L., & On, B. W. (2007). Are your citations clean? New scenarios and challenges in maintaining digital libraries. *Communication of the ACM, 50*(12), 33–38.

Lee, D., On, B. W., Kang, J., & Park, S. (2005). Effective and scalable solutions for mixed and split citation problems in digital libraries. *ACM SIGMOD Workshop on Information Quality in Information Systems*, 69–76.

Lu, Y., Nie, Z., Cheng, T., Gao, Y., & Wen, J. R. (2007). Name disambiguation using Web connection. In *Proceedings of AAAI-07 workshop on information integration on the Web*.

McRae-Spencer, D. M. & Shadbolt, N. R. (2006). Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation. In *Proceedings of the ACM/IEEE-CS joint conference on digital libraries* (pp. 53–54).

Osuna, E., Freund, R., & Girosi, F. (1997). ‘‘Support vector machines: training and applications", AI Memo 1602, Massachusetts Institute of Technology, 1997b.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H. F., Goncalves, M. A., & Ferreira, A. A. (2009). Using Web information for author name disambiguation. In *Proceedings of the ACM/IEEE-CS joint conference on digital libraries* (pp. 49–58).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 398–403.

Smith, D. A. & Crane, G. (2002). Disambiguating geographic names in a historical digital library. In *Proceedings of the European conference on digital libraries* (pp. 127–136).

Song, Y., Huang, J., Councill, I. G., Li, J. & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In *Proceedings of the ACM/IEEE joint conference on digital libraries* (pp. 342–351).

Tan, Y. F., Kan, M. Y. & Lee, D. (2006). Search engine driven author disambiguation. In *Proceedings of the ACM/IEEE joint conference on digital libraries* (pp. 314–315).

Treeratpituk, P. & Giles, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the ACM/IEEE-CS joint conference on digital libraries* (pp. 39–48).

Vu, Q. M., Takasu, A., & Adachi, J. (2008). Improving the performance of personal name disambiguation using Web directories. *Information Processing and Management, 44*(4), 1546–1561.

Yang, K. H., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2006). Extracting citation relationships from Web documents for author disambiguation. Technical Report (TR-IIS-06-017), Institute of Information Science, Academia Sinica.

Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). Author name disambiguation for citations using topic and Web correlations. In *Proceedings of the European conference on research and advanced technology for digital libraries* (pp. 14–19).