中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-06-017

Extracting Citation Relationships from Web Documents for Author Disambiguation

Kai-Hsiang Yang, Jian-Yi Jiang, Hahn-Ming Lee, Jan-Ming Ho

December 26, 2006 || Technical Report No. TR-IIS-06-017

http://www.iis.sinica.edu.tw/LIB/TechReport/tr2006/tr06.html

Extracting Citation Relationships from Web Documents for Author Disambiguation

Kai-Hsiang Yang[*], Jian-Yi Jiang [#], Hahn-Ming Lee[*,#], Jan-Ming Ho[*]

[*] Institute of Information Science,
Academia Sinica, Taipei,Taiwan
{khyang, hmlee, hoho}@iis.sinica.edu.tw

[#] Department of Computer Science and
Information Engineering National Taiwan
University of Science and Technology
{M9315026, hmlee} @mail.ntust.edu.tw

ABSTRACT

Disambiguating the citation records of authors with the same name is a very interesting and challenging problem that affects many research and application fields, such as digital libraries. However, current bibliographic digital libraries like CiteSeer can not correctly disambiguate citation records because of two problems: information sparsity (citations for an individual have few or no common features), and information noise (citations for different individuals have the same coauthor names, title words, or venue words). To resolve these problems, we propose a novel author disambiguation scheme that searches for authors' publication lists on the Web to enrich citation information. A binary classifier and a cluster separator are used to filter out noise. The experiment results show that the disambiguation accuracy improves from 51% to 73% when Web information is used in the disambiguation task. Furthermore, for most datasets, the clustering precision rate is satisfactory (more than 90%).

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; H.3.7 [**Digital Libraries**]: Dissemination

General Terms

Design, Experimentation.

Keywords

Citation record clustering, author disambiguation.

1. INTRODUCTION

Researchers often search for relevant works by using bibliographic digital libraries, such as DBLP, CiteSeer, and Pubmed Medline. However, at present, such libraries cannot correctly auto-index the citation records of authors. For example, Han et al. [11] found that the author page of "Yu Chen" in DBLP contains citations authored by three individuals with the same name. In CiteSeer, Oyama et al. [16] observed that the statistics of "Most Cited Authors in Computer Science" also have the same problem, especially when the name is abbreviated; for example, when it only contains the first initial and the last name. The above problems are caused by the name ambiguity problem, which means that multiple individuals use the same name label.

Since name labels that are the same are ambiguous, they are not very useful in the disambiguation of citation records. In fact, disambiguating authors with the same name label in citation records must address two problems: information sparsity and information noise. The former means that the available information is not sufficient to enable correct disambiguation, while the latter means that information about ambiguous names is not useful for disambiguation. For example, in Fig. 1, the "J. Smith" in citations (1) and (2) is the abbreviated name of the individual, "John R. Smith", and the "J. Smith" in citation (3) is the abbreviated name of the individual, "Jim Smith". However, citations (1) and (2) have no common features except the author name, "J. Smith". Therefore, the "J. Smith" in these two citations could be identified as different individuals if disambiguation is based on the citation text only. This problem is called information sparsity. On the other hand, citations (2) and (3) not only have the same author name "J. Smith", but they are also published by the same conference, "CIKM". Based on the citation text, the "J. Smith" of citations (2) and (3) could be identified as the same individual. This problem is called information noise.

(1) A. Natsev, Y. Chang, **J. Smith**, J. Vitter: Supporting Incremental Join Queries on Ranked Inputs. VLDB 2001:281-290

(2) **J. Smith**, C. Li: An Adaptive View Element Framework for Multi-dimensional Data Management. CIKM 1999:308-315

(3) **J. Smith**, S. Sampaio, P. Watson, N. Paton: Polar: An architecture for a parallel odmg compliant object database. CIKM 2000:352-359

Fig. 1. Examples of information sparsity and information noise

In this paper, we propose an author disambiguation scheme to address the information sparsity and the information noise problems. Our approach is based on collecting authors' publication lists on the Web as data to help in the disambiguation task. We also adopt a pair-wise clustering algorithm and introduce a cluster separator to reduce the effect of information noise on author disambiguation. According to our experiment results, the accuracy of author disambiguation increases from 51% to 73% by using the proposed approach. In addition, for most datasets, the clustering precision rate is satisfactory (more than 90%), i.e. most citations in the same cluster refer to the same individual.

The remainder of the paper is organized as follows. Section 2 reviews related works, and Section 3 describes the proposed disambiguation approach. In Section 4, we detail and discuss the

WWW 2007, May 8--12, 2007, Banff, Canada.

experiment results. Finally, in Section 5, we present our conclusions and indicate the direction of our future work.

2. RELATED WORK

A great deal of research has focused on the name ambiguity problem in different types of data, such as place name disambiguation [18], gene vs. protein name disambiguation [8], personal name disambiguation in documents [7, 14] and on web pages [1, 13, 15]. To resolve the information sparsity problem, some related information is used to facilitate the disambiguation task, for example, the concepts of words [8], hyperlink structures and social networks [1, 14, 15], and the statistics of search results from search engines [7]. For example, Han et al. [10] try to improve author disambiguation accuracy by clustering title words and venue words with similar concepts, while Tan et al. [19] perform author disambiguation by measuring the similarities between web documents related to citations.

In general, disambiguation methods can be categorized as supervised learning methods [9, 12] and unsupervised learning methods [1, 11, 13]. Because background information on each individual is not always available, it is difficult to disambiguate author citations with supervised methods. Thus, we believe that an unsupervised learning method is more suitable for the task.

A clustering algorithm based on a learned pairing function is often used in duplicate detection [2, 6], and has also been applied successfully to personal name disambiguation [3, 7, 16]. The algorithm generates a vector of similarity scores, computed by comparing the attributes of two records, and then uses a trained binary classifier to label each pair of records as matched or non-matched. A matched pair means that the two records refer to the same entity, whereas the records in a non-matched pair refer to two different authors. When clustering records by using matched pairs, all records referring to the same individual are clustered together. Therefore, background knowledge, such as the number of individuals with the same name or each individual's identification information, is not required by a pair-wise clustering algorithm. However, there is a problem with this kind of algorithm in that the inclusion of a few falsely matched pairs, i.e., non-matched pairs, substantially reduces the accuracy of the clustering result.

Our contribution in this paper is that we combine and analyze citation features and information derived from the Web. In addition, to avoid clustering errors generated by the pair-wise clustering algorithm, we propose a cluster separator based on graph structure detection.

3. PROPOSED APPROACH

First, we formally define the citation name identity problem as follows. Given a collection of citations, all of which contain identical author names, our goal is to cluster citations for the same author; i.e., citations in the same cluster that refer to the same author. The proposed disambiguation approach is described in following subsections.

3.1 Feature Generation

The first step of our approach extracts and generates representative features from citation texts. We use three familiar attributes of citations as known citation information: **coauthor**, **title,** and **venue**. It is assumed that the citation attributes can be extracted and identified by some methods, such as rule-based systems **[4]** or a hidden Markov model **[17]**. Because the properties of attributes are different, the features of each citation attribute are generated individually. Here, a feature could be a coauthor's name or a word. Note that the title words and venue words are pre-processed by stemming and stop-word elimination.

In addition, because works with ambiguous author names in an authors' publication list were probably written by the same individual, the accuracy of author disambiguation could be improved by utilizing this kind of information. We therefore utilize authors' publication lists on the Web to facilitate disambiguation. As the **title** is the representative attribute in a citation, we use it to query a search engine first. Then the URLs of web documents containing the phrase **title** of a citation are retrieved as candidates for the authors' publication lists. However, the authors' publication lists in some digital libraries, such as DBLP and Docis, may contain noisy information because, as mentioned earlier, some digital libraries cannot correctly auto-index the citation records of authors. Furthermore, because publication lists are not edited by the authors themselves, they may contain errors. To resolve this problem, we filter the URLs of web documents in digital libraries by a simple method, i.e., hostname matching and the keyword matching. If a URL contains a specific hostname or certain keywords, they are filtered out. Then, the remaining URLs are defined as the features of the **web** attribute for the citation.

The process of feature generation can be summarized as follows. A citation d is represented as a collection of four feature sets, $\{C_d, T_d, V_d, W_d\}$, where C_d, T_d, V_d, and W_d are the feature sets of **coauthor**, **title**, **venue**, and **web** respectively.

3.2 Pair-wise Clustering Algorithm

After generating all citation features, we apply a clustering algorithm based on a learned pairing function.

3.2.1 Generating Pair-wise Vectors

The pair-wise clustering algorithm first calculates the similarity scores between the corresponding attributes of any two citations, and then represents the scores as a vector. Because the properties of different attributes could be dissimilar, we use different types of similarity metrics for similarity calculation.

- **Similarity Metrics for Citation Attributes**

For the citation attributes, **coauthor, title, and venue**, similarity calculation is based on two disambiguation concepts. First, if the corresponding feature sets of two citations are similar, the two works were probably authored by the same individual. This would be the case if the titles of the two works are similar. Second, if there are several common features in the corresponding feature sets of two citations, the two works were also probably authored by the same individual. For example, it is more likely that two citations belong to the same individual if they have three or more common coauthor names. To measure the relative importance of the two disambiguation concepts, the following two similarity metrics are applied to the three citation attributes: the Cosine Similarity Metric (CSM) and the Modified Sigmoid Function (MSF).

➢ *Cosine Similarity Metric (CSM)*

The cosine similarity metric, also called the cosine distance function, is often used to estimate the similarity of strings.

Because two citations are probably authored by the same individual if they have similar coauthor names, title words or venue words, CSM can be used to disambiguate author citations. The cosine similarity score of two feature sets *X* and *Y, CSM*(*X*, *Y*), is calculated as follows.

$$CSM(X,Y) = \frac{\sum_{f \in X \cap Y} TFIDF(f,X) \cdot TFIDF(f,Y)}{\sqrt{\sum_{f \in X} TFIDF(f,X)^2} \cdot \sqrt{\sum_{f \in Y} TFIDF(f,Y)^2}} \quad (1)$$

where *f* is a feature in *X* or *Y*, *TFIDF*(*f* , *X*) is the TFIDF weight of *f* in *X*, and *TFIDF*(*f* , *Y*) is the TFIDF weight of *f* in *Y*. If a corresponding attribute of two citations has several similar or common features with high TFIDF weights, the cosine similarity score for that attribute will be closer to 1, which means that the two works were probably authored by the same individual.

- *Modified Sigmoid Function (MSF)*

To measure the second disambiguation concept, the similarity score is increased according to the number of common features in two feature sets. To do this, we modify and apply the sigmoid function. Given two feature sets, *X* and *Y*, the similarity score of the MSF, *MSF*(*X*,*Y*), is calculated as follows.

$$MSF(X,Y) = \begin{cases} 0 & \text{, if } |X \cap Y| = 0 \\ \dfrac{1}{1 + \mathrm{e}^{-(\alpha \cdot |X \cap Y| - 4)}} & \text{, otherwise} \end{cases} , \quad (2)$$

where $|X \cap Y|$ is the number of features at the intersection of *X* and *Y*, and α is a parameter used to adjust this function for different attributes. The value of α should be increased if citations authored by the same individual frequently have few identical attributes, such as coauthor; otherwise, it should be reduced. By applying the MSF, the similarity score of two citations will be closer to 1 when they have several identical features for the same attribute.

- **Similarity Metrics for Web Attribute**

For the web attribute, the main disambiguation concept is that web documents containing a large number of citations with the same author's name are usually authors' publication lists. To measure this disambiguation concept, we use the Maximum Normalized Document Frequency (MNDF), which is described below.

- *Maximum Normalized Document Frequency (MNDF)*

Because web features are URLs, which are unique, citations containing identical web features are included in the same web document. Consequently, authors' publication lists can be extracted by finding the features with the highest citation frequency at the intersection of any two citations' web feature sets. The citation frequency of a web feature is the number of citations containing that feature. Moreover, because the citation frequencies of web features are affected by the size of the citation set, we use the highest citation frequency for normalization. Given two web feature sets, *X* and *Y*, we calculate their MNDF similarity score, *MNDF*(*X*, *Y*), as follows:

$$MNDF(X,Y) = \begin{cases} 0 & \text{, if } X \cap Y = \varnothing \\ \dfrac{\max_{f \in X \cap Y} (\mathrm{DF}_f)}{\max_{\forall f} (\mathrm{DF}_f)} & \text{, otherwise} \end{cases} , \quad (3)$$

where DF_f is the number of citations that contain the web feature *f*, i.e., the citation frequency of *f*. If two citations have a common web feature with the number of citations close to the maximum citation frequency in the ambiguous citation set, their MNDF similarity score will be close to 1.

Pair-wise vector generation can be summarized as follows. Any two citations in the ambiguous citation set can be used to generate a 7-dimensional feature vector. If there are two ambiguous citations, $s = \{C_s, T_s, V_s, W_s\}$ and $t = \{C_t, T_t, V_t, W_t\}$, their pair-wise vector will be represented as follows: ($CSM(C_s,C_t)$, $MSF(C_s,C_t)$, $CSM(T_s,T_t)$, $MSF(T_s,T_t)$, $CSM(V_s,V_t)$, $MSF(V_s,V_t)$, $MNDF(W_s,W_t)$). Note that each feature value of the vector is in the range 0 to 1.

3.2.2 Clustering Citations by Labeling Citation Pairs

After generating the pair-wise vectors of any two citations, we can cluster the citations according to the relationships between all the vectors. Because the contribution of different attributes to disambiguation can be learned from the training set, the learned pairing function mitigates the effect of information noise. Moreover, to prevent clustering mistakes caused by labeling errors, we train the pairing function by increasing the penalty for falsely matched pairs in the training phase until the most accurate disambiguation cluster result is obtained.

Next, the pairs labeled as matched are used to build citation clusters. Here, the citations are clustered by constructing a graph in which a vertex represents a citation, and an edge represents a matched pair; that is, two vertices are connected if the pair of citations is labeled as matched. Then connected components in the graph are deemed citation clusters.

3.3 Cluster Separator

Although most information noise can be filtered out by the learned pairing function, some pairs will still be incorrectly matched because of information noise caused by attributes with high weights. Consequently, the disambiguation accuracy will be affected by these pairs. Interestingly, citations containing high levels of noise are often special cases, and the correct citation clusters can be merged by only one falsely matched pair. To deal with the problem of a high level of information noise, we propose a cluster separator based on graph structure detection. It filters out falsely matched pairs by removing the bridges in the graph.

However, many correctly matched pairs will also be filtered if all the bridges in the graph are removed. Because connected components in the graph are deemed citation clusters and citations in different clusters are identified as belonging to different authors, the disambiguation accuracy would be impaired if correctly matched pairs were filtered out. Hence, a threshold is set for choosing bridges that should be removed. Then, a bridge is removed if the numbers of vertices in the two divided connected components are both above the given threshold. After all the relevant bridges have been removed, the remaining connected components in the graph represent the disambiguation result.

4. EXPERIMENTS

In the experiments, we use the datasets constructed by Han et al. **[11]**, which contain citations collected from the DBLP computer science bibliography. Each citation consists of the three attributes discussed previously, namely, coauthor, title and venue. We had to change all abbreviated publication venues to their full venue names manually. In addition, all author names in the citations were reduced to the initial of the first name plus the last name because this format is common in citations. Then, citations with the same popular name, such as "A. Gupta", "C. Chen", "J. Smith" or "K. Tanaka" were grouped as a dataset. Han et al. selected 14 popular names to create datasets, and manually labeled the author citations in each dataset for evaluation, as shown in Table 1.

Table 1. The 14 DBLP Datasets. N denotes the number of individuals, C denotes the number of citations, (#C) indicates the range of the number of citations and #N is the number of individuals whose citations are within the range (#C).

	Name	N	C	Distribution of citations: (#C):#N
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
2	A. Kumar	14	244	(2~10):9, (11~20):1, (21~30):2, (41~50):1, (91~100):1
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
4	D. Johnson	15	368	(2~10):9, (11~20):1, (21~30):1, (31~40):3, (181~190):1
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
6	J. Martin	16	112	(2~10):12, (11~20):3, (21~30):1
7	J. Robinson	12	171	(2~10):6, (11~20):3, (21~30):2, (41~50):1
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
9	K. Tanaka	10	280	(2~10):5, (11~20):1, (31~40):2, (61~70):1, (101~110):1
10	M. Brown	13	153	(2~10):8, (11~20):2, (21~30):2, (41~50):1
11	M. Jones	13	259	(2~10):7, (11~20):1, (31~40):2, (41~50):2, (51~60):1
12	M. Miller	12	412	(2~10):7, (11~20):2, (21~30):1, (141~150):1, (191~200):1
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
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

4.1 Evaluation Method

Like Han et al. **[11]**, we evaluate the experiment results in terms of the disambiguation accuracy, calculated by dividing the sum of correctly clustered citations by the total number of citations in the dataset. Because citations in different clusters belong to different individuals, an author should have at most one correct citation cluster. To find the correct cluster of each individual, we first perform cluster assignment. A citation cluster is assigned to the author who has the most citations among the authors in that cluster. Each citation cluster assigned to an individual is a candidate for his/her correct cluster. Then, the correct cluster for an individual is the candidate cluster with the most citations. The disambiguation accuracy is then calculated as follows:

$$Accuracy = \frac{\sum_{i \in I} n_{ir}}{N}, \qquad \textbf{(4)}$$

where I is the set of individuals in the dataset, r is the correct cluster of individual i, and N is the total number of citations in the dataset.

Besides the disambiguation accuracy, we use two traditional evaluation methods, precision and recall, to determine the precision of the clustering result and the effect of attributes on author disambiguation. For each dataset, the clustering precision and clustering recall are calculated as follows:

$$Precision_{cluster} = \sum_{g \in G} \frac{n_g}{N} \cdot \frac{n_{ig}}{n_g}, \qquad \textbf{(5)}$$

$$Recall_{cluster} = \sum_{g \in G} \frac{n_g}{N} \cdot \frac{n_{ig}}{n_i}, \qquad \textbf{(6)}$$

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

4.2 Experiment Results

The goal of our experiment is twofold: performance evaluation and attribute analysis. The performance of our approach is evaluated in terms of the disambiguation accuracy, while attribute analysis determines the effect of the attributes and similarity metrics on author disambiguation. The experiment results are discussed in following subsections.

4.2.1 Performance Evaluation

First, we divided the 14 DBLP datasets into two parts because the pairing function needs training data. Datasets 1 to 7 were called Part I, and the others were called Part II. When one part was used for training, the other was used for testing. The pairing function was C-SVC with an RBF kernel, implemented by LibSVM **[5]**. The training parameters were set at $\boldsymbol{C}_{-+}$:$\boldsymbol{C}_{+-}$=1:4, γ=8 when Part I was used as training data, and at $\boldsymbol{C}_{-+}$:$\boldsymbol{C}_{+-}$=1:8, γ=8 when Part II was used. The threshold of the cluster separator was set at 5; and the α of MSF was set at 2 when MSF was applied to the coauthor attribute, and at 1 when it was applied to the title attribute or the venue attribute. We compared our disambiguation results with those of Han et al. **[11]**, as shown in Fig. 2. Because the web attribute with the MNDF metric cannot be used in Han's approach, only the results based on the three citation attributes are presented. The "K-way spectral clustering method" is Han et al's approach without using web attribute information. The "Proposed approach (without web info.)" refers to citations disambiguated by our approach without using web information, while the

"Proposed approach (with web info.)" means the web attribute with the MNDF metric was used.

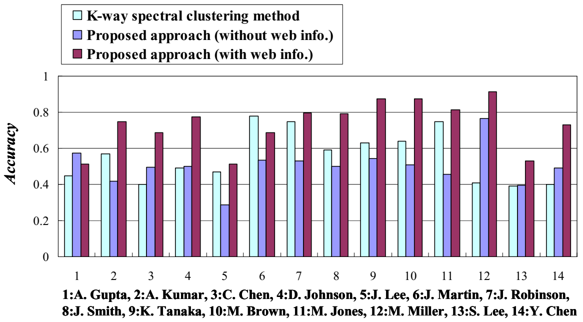

Fig. 2. Comparison of Han's K-way spectral clustering method and our approach

As shown in Fig. 2, the disambiguation accuracy for some datasets in our approach was better than Han's results when the web attribute with the MNDF metric was not used, especially in the four datasets "A. Gupta", "C. Chen", "M. Miller", and "Y. Chen". This demonstrates that our approach can mitigate the effect of information noise on author disambiguation. Even so, the disambiguation accuracy for several datasets was worse than that achieved by the K-way spectral clustering method. The reason is that the contribution of an attribute to disambiguation of different datasets may vary; in other words, information that is useful for some datasets may be lost when the pairing function is trained for general purposes.

When the web attribute with the MNDF metric is used, the disambiguation accuracy for most datasets improves substantially. Thus, information provided by the web attribute with the MNDF metric helps resolve the problem of information sparsity in author disambiguation. However, the disambiguation accuracy of the dataset "A. Gupta" is impaired because two individuals with the name "A. Gupta" coauthor the same papers. Consequently, many citations for the two individuals were clustered together when a publication list obtained from the Web was used for disambiguation.

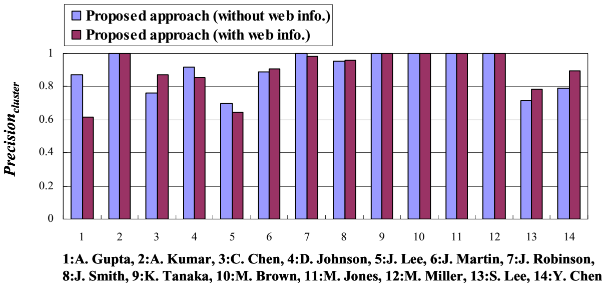

Fig. 3. The clustering precision of each dataset

We also evaluated the performance of the cluster separator on each dataset. The experiment results, listed in Table 2, show that the cluster separator improves the disambiguation accuracy when the web attribute with the MNDF metric is used, especially for the datasets "C. Chen", "J. Lee", "S. Lee", and "Y. Chen". In other words, the cluster separator is effective in removing falsely matched pairs. In contrast, the disambiguation accuracy of most datasets was impaired when the web attribute with the MNDF metric was not used because the cluster separator filtered out some correctly matched pairs from the datasets, as shown by the results. We think the threshold of the cluster size may need to be increased when the information sparsity problem is serious.

Table 2. The effect of using the cluster separator

		Without web info.		With web info.	
	Name	Without separator	With separator	Without separator	With separator
1	A. Gupta	0.529	0.572	0.51	0.513
2	A. Kumar	0.467	0.418	0.75	0.75
3	C. Chen	0.459	0.495	0.629	0.688
4	D. Johnson	0.5	0.5	0.774	0.774
5	J. Lee	0.242	0.286	0.486	0.513
6	J. Martin	0.5	0.536	0.688	0.688
7	J. Robinson	0.532	0.532	0.795	0.795
8	J. Smith	0.517	0.498	0.792	0.792
9	K. Tanaka	0.543	0.543	0.875	0.875
10	M. Brown	0.641	0.51	0.876	0.876
11	M. Jones	0.475	0.456	0.815	0.815
12	M. Miller	0.796	0.767	0.913	0.913
13	S. Lee	0.383	0.397	0.455	0.53
14	Y. Chen	0.558	0.493	0.685	0.733
Mean		0.510142857	0.500214286	0.717357143	0.7325

Finally, we calculated the precision of citation clusters in order to evaluate the confidence of our clustering results. From Fig. 3, we observe that the cluster precision of most datasets was high (≧ 0.8), even when the web attribute with the MNDF metric was used to improve the disambiguation accuracy. This means that most citations in the same cluster definitely belong to the same author. However, the clustering precision of the "A. Gupta" dataset was reduced substantially when the web attribute with the MNDF metric was used. The reason for this result is the same as in the special case mentioned above: two individuals with the same name coauthor the same paper(s). It is very difficult to disambiguate this kind of case correctly.

4.2.2 Attribute Analysis

In this experiment, we clustered the citations of the 14 DBLP datasets using multiple similarity thresholds to determine an attribute's similarity. That is, the citations were clustered by comparing the attribute's similarity with different thresholds. A citation pair was labeled as matched if its similarity score was higher than the given threshold. Note that all the similarity scores are in the range 0 to 1. Here, the setting of α is the same as in the previous experiment. The cluster separator was not applied in this experiment. The clustering precision and clustering recall of the 14 DBLP datasets are illustrated in Fig. 4.

As shown in the figure, the web attribute with the MNDF metric achieves a high cluster precision rate (≧0.9) when the cluster recall rate is lower than 0.5, which means the feature provides useful information with less noise for disambiguation. However, its maximum clustering recall is only about 0.75, which is probably due to exact keyword matching in search engines. Although the clustering recall rate can be improved by using partial keyword matching, many web documents unrelated to citations will also be retrieved. For this reason, the search scheme could be modified to match by partial phrases of titles to control the tradeoff between the number of retrieved web documents and their relevance to the citations. Of the three citation attributes, Fig. 4 shows that **coauthor** provides the most useful information for disambiguation, and **title** is slightly better than **venue**. However, the maximum cluster recall for the coauthor attribute is only 0.5, which suggests the information sparsity problem for the coauthor attribute is very serious. In addition, the similarity metric MSF for the three attributes achieves better clustering precision than the CSM metric. In other words, disambiguation information derived by the MSF metric contains less noise than that obtained by the CSM metric.

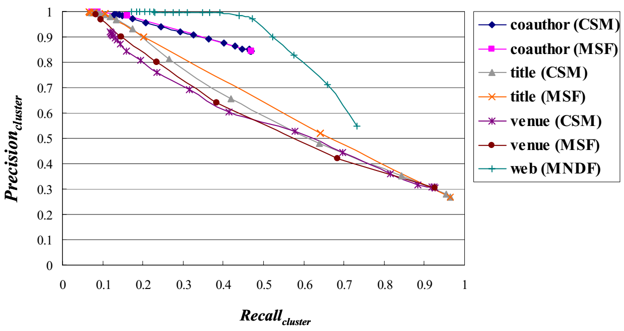

Fig. 4. The clustering precision and clustering recall of each dataset

5. CONCLUSION

We have addressed the problem of disambiguating citation records for different authors with the same name, and proposed a solution based on authors' publication lists downloaded from the Web. Our experiment results show that, when web information is used, the average disambiguation accuracy improves from 51% to 73%, while the average clustering precision rate is satisfactory (≧ 90%). Moreover, the disambiguation accuracy of some datasets can be further improved when a cluster separator is used. In summary, our approach not only clusters citations of the same author into the correct cluster more accurately, it also reduces the disambiguation errors in different individuals' citations grouped in the same cluster.

Our approach can also be applied to the name variation problem when an ambiguous citation set is constructed by some blocking methods. In this case, a large number of citations for individual can be helpful in retrieving the author's publication list.

Finally, although the experiment results show that the use of authors' publication lists from the Web is very effective in improving disambiguation accuracy, some issues still need to be addressed. For example, an author's citations are not always listed on his/her publication list, or the publication list may not be available on the Web. For this reason, in the future, we will try to extract more useful Web information, such as the statistics of search results [[7]], to disambiguate author citations more accurately.

6. REFERENCES

[1] Bekkerman, R., McCallum, A.: Disambiguating Web Appearances of People in a Social Network. *Proceedings of the International WWW Conference* (2005) 463–470.

[2] Bilenko, M., Mooney, R.: Adaptive Duplicate Detection Using Learnable String Similarity Measures. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003) 39–48.

[3] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S.: Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems*, (2003) 16–23.

[4] Califf, M., Mooney, R.: Relational learning of pattern-match rules for information extraction. *Proceedings of the international conference on Artificial Intelligence* (1999) 328–334.

[5] Chang, C.-C., Lin, C.-J.: LIBSVM : a library for support vector machines. (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Cohen, W., Richman, J.: Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002) 475–480.

[7] Fleischman, M., Hovy, E.: Multi-Document Person Name Resolution. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2004).

[8] Ginter, F., Boberg, J., Jarvinen, J., Salakoski, T.: New Techniques for Disambiguation in Natural Language and Their Application to Biological Text. *Journal of Machine Learning Research*, Vol. 5 (2004) 605–621.

[9] Han, H., Giles, C. L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two Supervised Learning Approaches for Name Disambiguation in Author Citations. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries* (2004) 296–305.

[10] Han, H., Xu, W., Zha, H., Giles, C. L.: A Hierarchical Naïve Bayes Mixture Model for Name Disambiguation in Author Citations. *Proceedings of the ACM symposium on Applied computing* (2005) 1065–1069.

[11] Han, H., Zha, H., Giles, C. L.: Name Disambiguation in Author Citations using a K-way Spectral Clustering Method. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries* (2005) 334–343.

[12] Koppel, M., Schler, J.: Authorship Verification as a One-Class Classification Problem. *Proceedings of the international conference on Machine Learning* (2004) 489–495.

[13] Lloyd, L., Bhagwan, V., Gruhl, D., Tomkins, A.: Disambiguation of References to Individuals. *IBM Research Report* (2005).

[14] Malin, B., Airoldi, E., Carley, K. M.: A Network Analysis Model for Disambiguation of Names in Lists. *Computational & Mathematical Organization Theory*, Vol. 11 (2005) 119–139.

[15] Mann, G. S., Yarowsky, D.: Unsupervised Personal Name Disambiguation. *Proceedings of the Conference on. Computational Natural Language Learning* (2003) 33–40.

[16] Oyama S., Manning, C. D.: Using Feature Conjunctions across Examples for Learning Pair-wise Classifiers. *Proceedings of European Conference on Machine Learning* (2004) 322–333.

[17] Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden Markov model structure for information extraction. *Proceedings of AAAI Workshop on Machine Learning for Information Extraction* (1999) 37–42.

[18] Smith, D. A., Crane, G.: Disambiguating Geographic Names in a Historical Digital Library. *Proceedings of European conference on digital libraries* (2002) 127–136.

[19] Tan, Y. F., Kan, M.-Y., Lee, D.: Search Engine Driven Author Disambiguation. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries* (2006) 314–315.