Author Name Disambiguation for Citations Using Topic and Web Correlation*

Kai-Hsiang Yang[1], Hsin-Tsung Peng[1], Jian-Yi Jiang[2], Hahn-Ming Lee[1,2], and Jan-Ming Ho[1]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
{khyang,m9115013,hmlee,hoho}@iis.sinica.edu.tw
[2] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
{M9315026,hmlee}@mail.ntust.edu.tw

Abstract. Today, bibliographic digital libraries play an important role in helping members of academic community search for novel research. In particular, author disambiguation for citations is a major problem during the data integration and cleaning process, since author names are usually very ambiguous. For solving this problem, we proposed two kinds of correlations between citations, namely, *Topic Correlation* and *Web Correlation*, to exploit relationships between citations, in order to identify whether two citations with the same author name refer to the same individual. The topic correlation measures the similarity between research topics of two citations; while the Web correlation measures the number of co-occurrence in web pages. We employ a pair-wise grouping algorithm to group citations into clusters. The results of experiments show that the disambiguation accuracy has great improvement when using topic correlation and Web correlation, and Web correlation provides stronger evidences about the authors of citations.

Keywords: Citation clustering, Citation analysis, Author disambiguation.

1 Introduction

Today, many digital libraries, such as DBLP and Citeseer, collect a large number of publication records, which are called "citations" in this paper, in order to provide a bibliography search service for academic community, since researchers often need to search for the latest work in their interests and use citation counts to measure the impact for a specific researcher. To have a consistent, accurate and up-to-date citation dataset is a very important task for all digital libraries. However, for various reasons, such as incomplete citation information or authors with the same name, digital libraries cannot always correctly map citations to authors [1, 2, 3]. For example, Han et al.

* This work was supported in part by the National Digital Archive Program (NDAP, Taiwan), the National Science Council of Taiwan under grants NSC 96-2628-E-011-084-MY3, NSC 96-2221-E-011-064-MY3, NSC 95-3114-P-001-002-Y02, NSC95-3114-P-001-001-Y02 and NSC 95-2221-E-001-021-MY3.

B. Christensen-Dalsgaard et al. (Eds.): ECDL 2008, LNCS 5173, pp. 185–196, 2008.

[1] found that the author page of "Yu Chen" contains citations authored by three individuals with the same name; Lee et al. [3] also found that the page of "Wei Wang" contains at least citations authored by four individuals. This problem is called the "name ambiguity problem", which means that multiple individuals share the same name. In recent years, name disambiguation has become a major challenge when integrating data from multiple sources in bibliographic digital libraries [3]. It needs more information to exploit the relationships between citations in order to improve the accuracy of name disambiguation [4].

In this paper, our contribution is to propose a novel solution for the name disambiguation problem by using two kinds of correlations between citations, namely, *Topic Correlation* and *Web Correlation*, to enrich insufficient citation information. For topic correlation, it is assumed that every researcher focuses on few research topics, and each of his/her publication is related to those topics. If two citations with the same author name have related topics, there is high probability that they belong to the same individual. To measure the topic relationships between citations, a topic association network is built by using the association rule mining technique [5]. We measure the strength of topic correlation between citations by the distances between them in the topic association network. For Web correlation, if two citations co-occur in a web page, such as the author's or co-author's publication list, they are probably related to the same researcher. The Web correlation is measured by the occurrence number of citations in the Web pages. After calculating these features, a pair-wise grouping algorithm is used to group the citations into clusters. Through extensive experiments, the average disambiguation accuracy for our system increases from 49% to 75% when both the topic correlation and Web correlation are used; in addition, both the average clustering precision and recall rates are satisfactory (more than 90% and 75%, respectively), i.e., most citations in a cluster refer to the same individual.

The remainder of the paper is organized as follows. Section 2 reviews the related work, and Section 3 describes the proposed disambiguation approach. In Section 4, we detail and discuss the 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 disambiguation problem in different types of data, such as geographic name disambiguation [6], biomedical term disambiguation [7], and personal name disambiguation [8]. Several papers [1, 9, 10, 11] have also focused on using the content in citations to solve the name disambiguation problem. However, the success of existing approaches has been limited due to insufficient information in the content of citations. To resolve this problem, some relational information is used to facilitate the disambiguation task. For example, Han et al. [12] try to improve disambiguation accuracy by clustering title words and venue words with similar concepts. Song et al. [13] introduce the relationships between authors and topics in citations to improve the disambiguation accuracy by extracting the word-based relationships for each topic. More recently, some work [4, 14, 15, 16, 17] tries to solve the problem by gathering Web pages related to citations.

In general, prior work can be categorized into supervised classification approaches [2, 10, 18, 19] and unsupervised clustering approaches [1, 13]. The supervised classification approaches try to model all authors' patterns from a set of training data, since the data usually provides insight into how to capture implicit domain knowledge, and this method can be quite accurate and reliable when the dataset is good enough. For the unsupervised approaches, ambiguous citations are clustered into groups of distinct authors by measuring the similarities between the attributes in the citations.

3 Proposed Approach

3.1 Topic Correlation

We consider two characteristics of the academic research. First, the publications' venue information in citations briefly represents the topics; for example, the full name of JCDL (the Joint Conference on Digital Libraries) covers several interesting topics, such as "information visualization" and "data mining", but its main topic is "digital libraries." Second, researchers (authors) often have specific research topics, so their publications should closely relate to their research topics; i.e., the topics of same author have the associated relationships among them.

Based on these facts, we extract citations' topics from venue information and discover the topic-based relationships to build a topic association network by using association rule mining technique [5]. We then measure the strength of topic correlation between two citations based on the distances between them in the topic association network. If the strength of the topic correlation is high enough, the citations may have higher probability to refer to the same person.

3.1.1 Topic Association Network

First, we extract the phrases from the venue of each citation by using a knowledge-based database in which each phrase is viewed as a topic of citation. A collection of topics of an author's citations is seen as a single transaction; for example, an author has two citations that belong to two topics, namely, "artificial intelligence" and "machine learning", and a transaction is {"artificial intelligence", "machine learning"}. We then look for frequent itemsets, i.e., groups of topics that commonly occur together.

We derive association rules from 2-itemsets whose support and confidence values are higher than the predefined thresholds by using the Apriori algorithm [20]. Here, the confidence value determines a primary-secondary relationship between topics. Since the confidence of a rule represents common items as consequents and rare items as antecedents, so we define the consequents as the main topics and antecedents as their sub-topics. The rules are grouped by constructing a directed graph in which a vertex represents a topic, an edge represents a primary-secondary relationship and the weight of an edge represents the confidence value of the rule; for example, we have two rules, {"machine learning"}=>{"artificial intelligence"} and {"natural language processing"}=>{"artificial intelligence"}, and can infer that "artificial intelligence" is the main topic, "machine learning" and "natural language processing" are its sub-topics.

Since the graph is highly connected, two topics are usually connected even if they are not related. We apply hMETIS [21], a *k*-way hypergraph partition algorithm, to split the hypergraph into several clusters in which the topics are closely related. We

call these clusters a *topic association network*. The strength of the topic correlation between citations is based on the distances between the citations in the network. If the distance between the topics of two citations is small, the topics are related; therefore, the citations may belong to the same author.

3.2 Web Correlation

For the Web Correlation, our basic assumption is that researchers' citations are usually listed in their publication lists or even listed in their co-authors' publication lists. Based on this assumption, if two citations occur in the same Web page, it shows high probability to belong to the same individual. Hence, we use the co-occurrence times for two citations in Web pages as the Web Correlation.

As a paper title is essential for a citation, we use each title to query a search engine, and then retrieve all the URLs of Web pages as candidates for the publication lists (more detailed results are in the paper [17]). However, to collect the publication lists edited by humans only, we filter the URLs of several digital libraries. The remaining URLs are taken as the valid data source for Web correlation. If two citations appear in the same URL, we use them as an instance of Web correlation.

3.3 Pair-Wise Grouping Algorithm

The pair-wise grouping algorithm includes a pair-wise similarity measure, a binary classifier, and a cluster filter. The steps of the algorithm are as follows: (1) Generate pairs of citations by using similarity metrics. (2) Use the training data to train a binary classifier. (3) Apply the classifier to determine whether the pairs are matched. (4) Combine the predicted results to group the citations into appropriate clusters. (5) Filter out the pairs that would cause the clusters sparse.

3.3.1 Pair-Wise Similarity Metrics

A citation is represented as a collection of five attributes, i.e., coauthor, title, venue, topic, and Web attributes. The pair-wise grouping algorithm calculates the similarity scores between the corresponding attributes of any two citations by using different types of similarity metrics.

3.3.1.1 Similarity Metrics for Coauthor, Title, and Venue. For the three attributes, *coauthor*, *title* and *venue*, we propose two similarity metrics. The details of each similarity metric are as follows.

- *Cosine Similarity Metric (CSM)*

The cosine similarity metric, also called the cosine distance function, is used to estimate the similarity between two vectors (or attributes). It is very suitable to calculate the similarity for the paper title attribute, because each title can be treated as a vector of words. The cosine similarity score of two attributes *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 (term frequency-inverse document frequency) 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)*

The CSM, however, may not be able to retrieve an important feature when the frequency of the feature is low; for example, the TFIDF method cannot correctly measure the similarity for the field of coauthors' names. To resolve the problem, we propose the MSF metric, which is based on the co-occurrences of features in two corresponding feature sets. When the number of common features in two feature sets is increased, the similarity score will be increased exponentially. Given two attributes, X and Y, the similarity score $MSF(X, Y)$ is calculated as follows.

$$MSF(X,Y) = \begin{cases} \dfrac{1}{1+e^{-(|X\cap Y|-\alpha)}} & if\ |X \cap Y| \neq \phi\ , \\ 0 & otherwise \end{cases} \tag{2}$$

where $|X \cap Y|$ is the number of features at the intersection of X and Y. The shift value α is a parameter used to adjust the MSF metric for different attributes and should be decreased if citations authored by the same individual frequently have few identical features, such as coauthors; Otherwise, it should be increased. 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.

3.3.1.2 Similarity Metrics for Topic Correlation. Topic correlation is based on the concept that if the topics of two citations are related, the citations probably refer to the same individual. We use the Topic Similarity Metric (TSM) to model our concept.

- *Topic Similarity Metric (TSM)*

As mentioned previously, we build a topic association network to model the relationships between topics. Two citations may have an associated relationship in terms of their topics, even though their venue attributes yield low similarity scores based on CSM and MSF. The similarity score of two topics X and Y, $TSM(X, Y)$, is calculated as follows.

$$TSM(X,Y) = \begin{cases} 1 - \dfrac{w(X,Y)}{\{max_{a,b\in G}\ w(a,b)\}+1} & if\ X,Y \in G\ , \\ 0 & otherwise \end{cases} \tag{3}$$

where G is the topic association network, and $w(X,Y)$ is the sum of edge weights from topic X to Y, or vice versa, in G. The maximum sum of weights between any two topics in G is used for normalization. To avoid getting 0 as the TSM metric, we add the denominator by 1. If $w(X,Y)$ is small, which means the two topics are close in the network, their TSM similarity score will be close to 1.

3.3.1.3 Similarity Metrics for Web Correlation. Web correlation is based on the concept that if citations co-appear in the same web pages many times, they are probably authored by the same researcher. To measure this concept, we use the Maximum Normalized Document Frequency (MNDF), which is described below.

- *Maximum Normalized Document Frequency (MNDF)*

Because citations containing identical URLs are included in the same Web pages, authors' publication lists can be identified by finding the URLs with the highest citation frequency at the intersection of any two citations' Web attributes. Given two Web attributes, X and Y, we calculate their MNDF similarity score, $MNDF(X, Y)$, as follows.

$$MNDF(X,Y) = \begin{cases} \dfrac{\underset{f \in X \cap Y}{Max}(DF_f)}{\underset{\forall f}{Max}(DF_f)} & if\ X \cap Y \neq \phi \\ 0 & otherwise \end{cases}, \tag{4}$$

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

3.3.2 Binary Classifier

After generating the pair-wise vectors of any two citations, we adopt a supervised learning method to capture authors' writing patterns and distinction between different authors' citations. Specifically, a binary classifier is used to learn the distribution of pair-wise vectors. Moreover, to deal with the unbalanced data problem, it should be trained by increasing the penalty for falsely matched pairs in the training phase until the most accurate disambiguation result is obtained.

Next, the pairs predicted as matched are used to build citation clusters. The citations are clustered by constructing an undirected 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 predicted as matched. Connected components in the graph are deemed citation clusters and citations in different clusters are identified as belonging to different authors.

3.3.3 Cluster Filter

Due to the impact of boundary errors caused by the binary classifier, a falsely matched pairs could merge clusters into one large cluster in the graph and thereby affects the final result. To deal with this problem, we propose using a cluster filter based on graph structure detection. The citations would be connected densely by filtering out the bridges in the graph.

In the cluster filter, a threshold is set for choosing which bridges should be removed. Then, a bridge is removed if the numbers of vertices in two separate, but connected, components are above the given threshold. After all the relevant bridges

have been removed, the remaining citations are connected more densely in the clusters and represent the disambiguation result.

4 Experiments

4.1 Experiment Setting and Evaluation Method

In our experiments, we use the dataset constructed by Han et al. [1], which contain the citations collected from the DBLP Website. Each citation consists of the three basic attributes discussed previously, namely, coauthor, title and venue. Han et al. selected 14 popular author names to create their dataset and manually labeled the citations in each author name for evaluation. We select the authors who have at least 2 citations as the dataset, where there are 476 individual authors and 8,441 citations. To increase the complexity of this problem as Han's work did [1], all author names were reduced to the initial of the first name plus the last name. In addition, the title words and venue words are pre-processed by stemming and stop-word elimination. For the details, please refer to [1].

To construct the topic association network, we discover the topic-based relationships in the dataset. Due to the small numbers of transactions in our dataset (there are 476 authors in our dataset, but the total numbers of authors in DBLP are approximately 468,000), association rules with low support and confidence values are discovered for the most part. We set the support threshold s at 3 and the confidence threshold c at 0 for retaining most of information after observing several experimental results. There are 209 topics in total, and we identify four main topics in the graph, namely, "architecture and networking", "artificial intelligence", "multimedia", and "information retrieval"; therefore, we set the k value of hMETIS at 4 in the topic association network. Moreover, to measure Web correlation, we use each citation's title to query Google's search engine in order to collect authors' publication lists.

To deal with the unbalanced data problem, we adopted the C-SVC binary classifier with an RBF kernel function, implemented by LibSVM[1], which is the weighted SVM for unbalanced data. We divided the dataset into two parts because we needed training data for the binary classifier. The data of author names, which are from "A. Gupta" to "J. Robinson", were called Part I, and the others were called Part II. When one part was used for training, the other was used for testing. To define the appropriate parameters for the binary classifier and cluster filter, we followed a grid method to scan all sets of parameters. The SVM parameters were set as $\boldsymbol{C}_{-+}$:$\boldsymbol{C}_{+-}$=1:4, γ=8 when Part I was used as training data, and as $\boldsymbol{C}_{-+}$:$\boldsymbol{C}_{+-}$=1:8, γ=8 when Part II was used. Moreover, the shift value α in the MSF metric was set at 4 for three attributes and the threshold of the cluster filter was set at 5 after observing several experimental results.

We evaluate the experiment results in terms of the *disambiguation accuracy*, which is calculated by dividing the sum of correctly clustered citations by the total number of citations in the dataset [1]. Besides the disambiguation accuracy, we use two traditional evaluation methods, namely, *precision rate* and *recall rate* (as used in [16]), to represent the effect of the clustering result and the effect of attributes on author disambiguation.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm

4.2 Experiment Results

4.2.1 Performance Evaluation

We compared our disambiguation results with those of Han et al. [1], as shown in Fig. 1. The results show that the disambiguation accuracy for some author names in our approach was better than that achieved by Han's method when both topic correlation and Web correlation were not used (only using three basic attributes), especially in the four author names: "A. Gupta", "C. Chen", "M. Miller", and "Y. Chen". Even so, the disambiguation accuracy for several author names was worse than that achieved by the *K*-way spectral clustering method. The reason is because the impact of each attribute varies for different author name, but the binary classifier is trained for all cases, which may cause some variance for different author names.

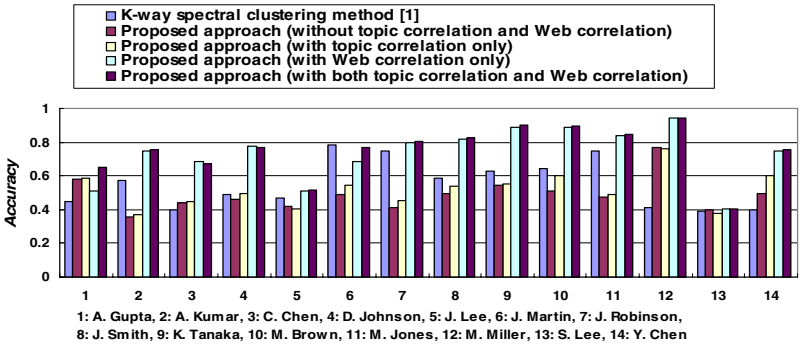

Fig. 1. Comparison of Han's *K*-way spectral clustering method and our approach

We can easily see that when only using topic correlation, the disambiguation accuracy was higher than both the Han's result and the result without using either correlation, even though only some author names, "A. Gupta", "C. Chen", and "Y. Chen", showed great improvements. It is reasonable to ask why topic correlation did not yield the obvious improvements in the experiment. The major reason is that, to discover the associations among topics by using association rule mining, a large number of transactions are needed. There are two issues should be discussed. First, due to the small number of transactions in the dataset, many related topics and relationships cannot be discovered. For example, two topics, namely, "digital library" and "knowledge management", should have a close relationship in the topic association network; however, this relationship was filtered because the number of transactions that include two topics is lower than the predefined support threshold. Therefore, the edge between "digital library" and "knowledge management" was not listed in the topic association network; besides, the topic "knowledge management" did not even listed in the network because all support values of the association rules that included it are lower than our threshold. Second, for retaining most of information, we set the support and confidence thresholds with low values. It may cause that the incorrect edges existed in the topic association network. For example, two topics "vldb" and "software engineering"

should not have an edge in the topic association network; however, in the topic association network, there was an edge between them, which has low support value 3 and low confidence value 0.167. Therefore, several incorrect edges still existed in the topic association network, and several pairs of citations led to incorrect measurement of the topic correlation. Besides, only 53.5% of the citations had topics listed in the network and can use the topic correlation information for author disambiguation as a result. These are the limitations of our experiments.

On the other hand, when only Web correlation was used, the disambiguation accuracy for most author names improved substantially, so the information provided by Web correlation helped resolve the disambiguation problem. Although the proposed approach with Web correlation performed well, there are some unexpected results in the dataset; for example, the disambiguation accuracy of the author name "A. Gupta" is impaired because two individuals with the name "A. Gupta" coauthored the same papers. Consequently, many citations for the two individuals were clustered together.

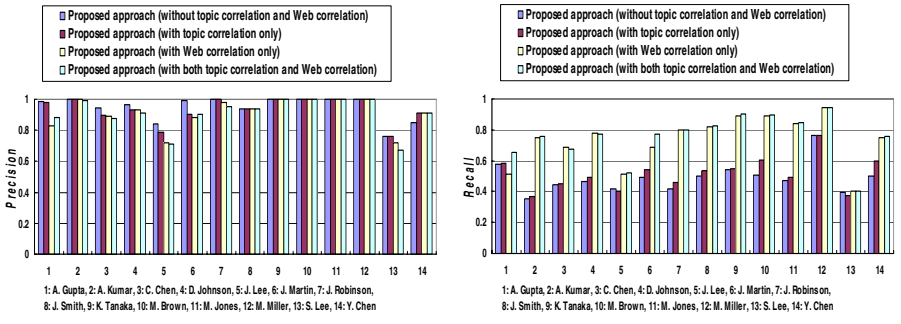

Fig. 2. The clustering precision and recall rates of our approach

When both correlations are used, we can see that Web correlation plays an important role in the author disambiguation task. Topic correlation also provides associations among citations, even though their venues have lower similarity measures based on string-based methods; for example, the accuracy of the author names "A. Gupta" and "J. Martin", was 51.3% and 68.8% when only Web correlation was used. However, by leveraging topic correlation, the accuracy of each dataset improves substantially (65% and 76.8%, respectively). In summary, the average disambiguation accuracy (75%) is higher than that without topic and Web correlations (49%) and Han's result (55% approximately).

We also calculated the precision and recall rates to evaluate our clustering results. As shown in Fig. 2, the clustering precision rates of most author names were high when both topic correlation and Web correlation were used. This means that most citations in the same cluster definitely belong to the same author. We also observe that both topic correlation and Web correlation enhance the clustering recall rates, which means an author's citations would be grouped into the same cluster, not separated into several clusters. The results show that Web correlation improves the recall rates markedly, but topic correlation does not. As mentioned previously, the major reason is the limitations of the experiments on topic correlation.

4.2.2 Attribute Analysis

We clustered the citations of our dataset using multiple similarity thresholds to determine an attribute's similarity. In other words, the binary classifier and cluster filter were not applied in this experiment. 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. The ROC curves of the dataset are illustrated in Fig. 3.

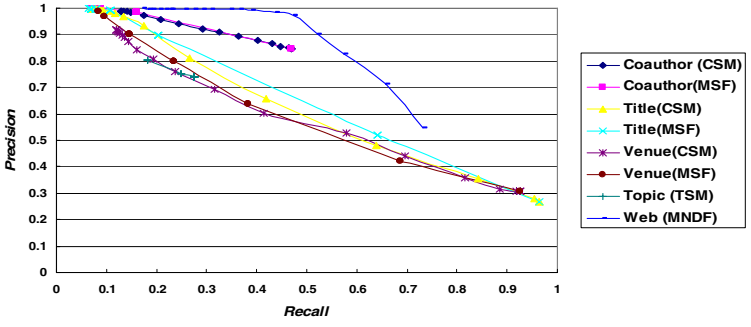

Fig. 3. The ROC curves for each attribute

We observe that topic correlation yields higher clustering precision rates ($\geqq 0.7$) and lower clustering recall rates ($\leqq 0.5$) when we set different similarity thresholds. As mentioned previously, only 53.5% of citations could be connected in the graph if TSM similarity scores of the pairs of these citations were above the threshold. The remaining citations did not have topics in the topic association network, so they did not have edges to connect with other topics in the graph; therefore, the precision rate was enhanced, but the improvement in the recall rate was limited. Besides, because the graph has the unidirectional and transitivity properties in this experiment, the citations belonging to two sub-topics were connected with each other even though two topics are not related to each other in the topic association network; for example, in a topic association network, we may find a topic "information process" has two sub-topics, "medical informatics" and "public key cryptography", which are not related to each other. However, in this experiment, the citations belonging to two sub-topics were connected with each other. Therefore, the precision and recall rates were influenced. That also explains why topic correlation did not work well.

The results also show that Web correlation achieves a high clustering precision rate ($\geqq 0.9$) when the clustering recall rate is lower than 0.5, which means the feature provides useful information with less noise for disambiguation. Of the three basic attributes, the coauthor attribute provides the most useful information for disambiguation, and title is slightly better than venue. In addition, disambiguation information derived by the MSF metric contains less noise than that obtained by the CSM metric.

5 Conclusion

We have addressed the problem of disambiguating citations for different authors with the same name. To solve the problem, we use additional information to exploit the

relationships between citations. We discover the implied topic-based relationships in citations to leverage name disambiguation, and show that the accuracy of disambiguation can be improved significantly by considering the publication lists on the Web. Our experiment results show that the average disambiguation accuracy improves from 49% to 75%, and both average precision and recall rates are satisfactory. In summary, our contribution is to group citations of the same author into the correct cluster more accurately, and proposes a useful solution for name disambiguation improvements.

In the future, we plan to enhance the approach in the following two directions: First, we will find out more transactions about research topics from other existent larger datasets, such as DBLP, ACM Digital Library and Citeseer. Second, we now only consider the Web pages edited by humans to measure Web correlation without using other existent digital libraries. In practice, an author's citations are not always listed on his/her publication list, or the publication list may not be available on the Web. In a future work, we will modify the way to measure the Web correlation by scaling other Web resources to disambiguate author citations more accurately. Furthermore, we will practically apply this approach to deal with the disambiguation problem in real world.

References

1. Han, H., Zha, H., Giles, C.L.: Name disambiguation in author citations using a K-way spectral clustering method. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 334–343 (2005)
2. Oyama, S., Manning, C.D.: Using Feature Conjunctions across Examples for Learning Pairwise Classifiers. In: Proceedings of European Conference on Machine Learning, pp. 322–333 (2004)
3. Lee, D., Kang, J., Mitra, P., Giles, C.L., On, B.W.: Are Your Citations Clean? New scenarios and challenges in maintaining digital libraries. Communication of the ACM 50(12), 33–38 (2007)
4. Lu, Y., Nie, Z., Cheng, T., Gao, Y., Wen, J.R.: Name Disambiguation Using Web Connection. In: Proceedings of AAAI 2007 Workshop on Information Integration on the Web (2007)
5. Han, E.H., Karypis, G., Kumar, V., Mobasher, B.: Clustering Based On Association Rule Hypergraphs. In: Proceedings of ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (1997)
6. Smith, D.A., Crane, G.: Disambiguating Geographic Names in a Historical Digital Library. In: Proceedings of European conference on digital libraries, pp. 127–136 (2002)
7. Al-Mubaid, H., Chen, P.: Biomedical Term Disambiguation: An Approach to Gene-Protein Name Disambiguation. In: Proceedings of the International Conference of Information Theory: New Generations, pp. 606–612 (2006)
8. Vu, Q.M., Masada, T., Takasu, A., Adachi, J.: Using a Knowledge Base to Disambiguate Personal Name in Web Search Results. In: Proceedings of the ACM symposium on Applied Computing, pp. 839–843 (2007)
9. Lee, D., On, B.W., Kang, J., Park, S.: Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries. In: Proceedings of ACM SIGMOD Workshop on Information Quality in Information Systems, pp. 69–76 (2005)

10. Han, H., Giles, C.L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two Supervised Learning Approaches for Name Disambiguation in Author Citations. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 296–305 (2005)
11. McRae-Spencer, D.M., Shadbolt, N.R.: Also By The Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 53–54 (2006)
12. Han, H., Xu, W., Zha, H., Giles, C.L.: A Hierarchical Naïve Bayes Mixture Model for Name Disambiguation in Author Citations. In: Proceedings of the ACM symposium on Applied Computing, pp. 1065–1069 (2005)
13. Song, Y., Huang, J., Councill, I.G., Li, J., Giles, C.L.: Efficient Topic-based Unsupervised Name Disambiguation. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 342–351 (2007)
14. Tan, Y.F., Kan, M.Y., Lee, D.: Search Engine Driven Author Disambiguation. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 314–315 (2006)
15. Kanani, P., McCallum, A.: 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 (2007)
16. Yang, K.H., Jiang, J.Y., Lee, H.M., Ho, J.M.: Extracting Citation Relationships from Web Documents for Author Disambiguation. Technical Report (TR-IIS-06-017), Institute of Information Science, Academia Sinica (2006)
17. Yang, K.H., Chung, J.M., Ho, J.M.: PLF: A Publication List Web Page Finder for Researchers. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pp. 295–298 (2007)
18. Culotta, A., Kanani, P., Hall, R., Wick, M., McCallum, A.: Author Disambiguation using Error-driven Machine Learning with a Ranking Loss Function. In: Proceedings of AAAI 2007 Workshop on Information Integration on the Web (2007)
19. Huang, J., Ertekin, S., Giles, C.L.: Efficient name disambiguation for large-scale database. In: Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 536–544 (2006)
20. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the International Conference on Very Large Data Bases, pp. 487–499 (1994)
21. Karypis, G., Aggarwal, R., Kumar, V., Shekhar, S.: Multilevel Hypergraph Partitioning: Applications in VLSI Domain. IEEE Transactions on VLSI Systems 7(1), 69–79 (1999)