A Study on Author Disambiguation based on Citation Data

Chi-Nan Hsieh and Kuang-hua Chen*
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan.
Email: r97126004@ntu.edu.tw, khchen@ntu.edu.tw

Abstract

In order to work out name disambiguation when retrieving information, researches on author identification are indispensable. This study attempts to address this problem by using citations in two out of 14 datasets from DBLP database with consideration of two additional features: year (Y) and page number (P), which have never been used before. Both supervised learning methods (Naïve Bayes and Support Vector Machine) and unsupervised learning method (K-means clustering) were employed to explore 28 different feature combinations. The findings show that the inclusion of feature Y improved the accuracy ratio, but no significant effects were observed when feature P was included. It is believed that experiments on the total 14 datasets will help us reach concrete conclusions of performance of different feature combinations.

Introduction

Names help identify a person with great ease. Yet, with widespread use of the Internet and digital information, name ambiguity problems have commonly occurred nowadays. The issues of name ambiguity include names in their abbreviated forms, typos, misspellings, multiple authors sharing the same name, and one author with multiple name labels. These often result in problems to researchers examining retrieval results of bibliographical databases. Name ambiguity affects not only the speed of information gathering but the consequent retrieval results. Many authorities are making their way towards the problem. Take International Standard Organization (ISO) for example. Their establishment of International Standard Name Identifier (ISNI) has been under way. In addition, universities and research institutions in the Netherlands have adopted Digital Author Identifier (DAI) in preparation for the coming of ISNI. Although the standard will be taken effect in the near future, lots of bibliographic documents and information with name ambiguities still need to be coped with. On the other hand, world renowned database vendors also contribute to solutions to the pressing problem. Elsevier (2009), for instance, provides "author search" function for its Scopus Database, and Thomson Reuters offers Distinct Author Identification System. These supplementary identification functions, however, still require complete author information to produce effect. Libraries usually build authority files in response to these ambiguities, such as OCLC WorldCat Identity Service and the Scholar Universe of ProQuest (2009). These name searches of identification mechanisms might achieve desired retrieval results, but they cannot handle a large amount of existent literature in databases without a lot of time and manpower.

Employing citations for author disambiguation is commonly found in this field, and some outside resources are taken into account for delivering satisfactory performance, such as employing parts of the full-text article (Song et al., 2007) and extracting information on web pages (e.g. Kanani et al., 2007; Yang et al, 2007, 2008). Nevertheless, copyright on full-texts and privacy concerns of author information can prove a hindrance to obtaining these supplementary resources. For these reasons, author disambiguation on "citation data" is considered feasible and reasonable. Related works on this area can be found in three studies by Han et al. (2004, 2005). In those studies, supervised methods and unsupervised methods are used for clustering and classification. The former can achieve accuracy ratio of 70%, and the latter 65%. However, only co-author names, article titles, and journal titles are used in Han et al. (2004, 2005) The purpose of this research is to explore performance of various feature combinations using complete citation data and investigate influences of additional features, such as "year" and "page number", on disambiguation.

Research Design

The dataset employed in this study was constructed by Han et al. (2005), which contains citations collected from DBLP database. The datasets consist of 14 popular author names, which include 476 individual authors and 8,441 citations. In order to increase the complexity of ambiguity, the first names of author names were changed into initials in Han's research design. However, this pilot study will use two out of 14 datasets and report the preliminary results.

* Corresponding author

The purpose of this study focuses performance of complete combinations of various features (e.g. authors, article titles, journal titles, venues) in citation data for disambiguation, although previous literature pointed out that the inclusion of all features at the same time might not necessarily achieve the best performance. Accordingly 28 feature combinations were explored in the study to examine how each kind of feature combination takes its effect. The framework is composed of three commonly used features Co-author (C), Article title (T), and Journal title (J) in combination with two additional features Year (Y) and Page number (P). The possible combinations are shown in Table 1.

Porter's stemmer was used for titles and journal titles in preprocessing data. In addition, stop words were removed with leaving keywords only. Since feature weighting is commonly concerned, TFIDF (Term Frequency and Inverse Document Frequency) scheme was adopted. Both supervised learning method and unsupervised learning method were employed to examine the performance. Two supervised learning methods were used, which are Naïve Bayes (Tool in NLTK) and LIBSVM for Support Vector Machine (Chang & Lin, 2010). The ratio of training set and testing set is 7:3, and cross validation was used. For unsupervised learning method, K-means clustering was conducted with SPSS software. And Euclidean distance was used as the measure of clustering distance between feature vectors. The performance is evaluated by dividing the sum of correctly clustered citations by the total number of citations in the dataset, i.e., commonly known as accuracy.

In order to consider features Year (Y) and Page number (P) in the study, year and page number in citation data have to be transformed into corresponding codes. For Year (Y) feature, it is assumed that each author has his/her period of academic production, so age distribution of the whole dataset was segmented into intervals. According to the dataset, the years of literature in DBLP mainly fall during year 1980 to 2005. Because of the time span, three age distributions were applied with 8 years as an interval. When it comes to Page number (P), under the influence of publication types and author preference, page numbers of the citation data were calculated first and intervals were set based on page number restrictions of publications. Three segmented points were designed in the study: 2 pages for poster papers, 8 pagers for conference papers, and more than 17 pages for journal papers. Then four intervals were constructed: fewer than 3 pages, 3 to 8 pages, 9 to 17 pages, and more than 17 pagers. In addition to the four intervals, two cases are considered: no page number and one page. Therefore, totally six cases for page number were considered. The research procedure is shown in Figure 1.

Preliminary Results

In this pilot study, 2 datasets out of 14 DBLP have been examined, which are J. Martin and A. Gupta. The J Martin Dataset (JM Dataset) includes 15 authors sharing the same name and 103 citations, and the A Gupta Dataset (AG Dataset) has 26 authors sharing the same name and 572 citations. Each feature and the feature combinations are discussed, and the influence of additional features Y and P are examined as follows.

Author disambiguation without considering additional features are presented in Figure 2 and 3 below. With single-feature experiment, performances of feature J of JM Dataset are 38.8% (K-means), 47.0% (Naïve Bayes), and 56.3% (SVM), respectively. The performance of feature C is less satisfactory, and that of feature T is least desired. In contrast, no obvious difference in performance for AG Dataset was found. With two-feature experiment, feature CJ achieved the most desired performance: 62.3% (SVM). The performance of feature TJ is less satisfactory, and that of feature CT is least desired. It is also noted that no prominent feature combinations were observed in the AG Dataset. With three-feature experiment, the JM Dataset achieved better performance (56.3%) when K-means clustering was used. When features C, T, and J were used for disambiguation at the same time, the combination cannot necessarily ensure best performance.

The findings of considering features Y and P were described as follows. With the inclusion of feature Y (see Figures 4 and 5), the average growth rates of JM Dataset is 2.40% (*sd = 3.93*) and that of AG Dataset is 0.13% (*sd = 2.06*). After adding feature P to the disambiguation, the average growth rates of JM Dataset is 0.22% (*sd = 3.61*), but that of AG Dataset is -0.62% (*sd = 1.46*). Finally, when features Y and P were included at the same time (see Figures 6 and 7), the average growth rates of JM Dataset is 2.05% (*sd = 4.48*) and that of AG Dataset is -0.36% (*sd = 2.25*). From the findings above, it is shown that feature Y improved performance substantially in the JM Dataset, while the inclusion of feature P produced negative effect. Furthermore, the performance of AG Dataset was interfered due to noise when feature Y (0.13%), feature P (-0.62%), and feature YP (-0.36%) were included.

Conclusions

As far as the findings of the two datasets are concerned, it is found that the inclusion of features exhibits different influence on disambiguation. The inclusion of features J and Y in the JM Dataset benefit the disambiguation considerably, but no significant effects on disambiguation is observed in the AG Dataset. When the two datasets were experimented with Expectation Maximization (EM) through hierarchical Naïve Bayes mixture model (Han, 2005), the JM Dataset could achieve accuracy of 80.3%, and the AG Dataset 56.2%. This difference shows data complexity of the two datasets. Apart from the datasets themselves, features of co-authors, article titles, and journal titles still produce stable disambiguation effect. Whether the additional features like Year and page number can greatly improve performance of the 14 datasets still require further experiments. At present, the inclusion of feature Y is positive and potential, while that of feature P still needs further investigation. Future research requires the completion of the rest 12 datasets so that concrete findings can be made and hypotheses can be verified. Moreover, the intervals of year of publications and pager numbers could be adjusted for careful exploration.

References

Chang, C.-C. and Lin, C.-J. (2010). *LIBSVM - A Library for support Vector Machines (Version 3.0)*. Retrieved Oct. 4, 2010, from http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Digital Author Identifier (DAI). (2009). *DAI-Standard wiki*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://www.surffoundation.nl/wiki/display/standards/DAI

Han, H., Giles, L., Zha, H., (2005). Name disambiguation in author citations using a K-way spectral clustering method. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.9354&rep=rep1&type=pdf

Han, H., Giles, L., Zha, H., Li, C., Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://clgiles.ist.psu.edu/papers/JCDL-2004-author-disambiguation.pdf

Han, H., Giles, L., Zha, H., Xu, W. (2005). A hierarchical Naïve Bayes mixture model for name disambiguation in author citations. In *Proceedings of the 2005 ACM symposium*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://clgiles.ist.psu.edu/papers/SAC-2005-Naïve-Bayes-Mixture.pdf

International Standard Name Identifier (ISNI). (2009). *ISNI Draft ISO 27729*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://www.isni.org/

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource bounded information gathering from the web. In M. M. Veloso (Ed.), *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 429-434.

OCLC. (2009). *WorldCat Identity Service*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://orlabs.oclc.org/identities

ProQuest. (2009). *Scholar Universe*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://www.scholaruniverse.com

Song, Y., Huang, J., Councill, I. G., Li, J. & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In E. M. Rasmussen, R. R. Larson, E. Toms, S. Sugimoto (Eds.), *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 342-351.

Thomson Reuter. (2009). *Distinct Author Identification System*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://scientific.thomsonreuters.com/support/faq/wok3new/dais/

Yang, K. H., Jiang, J. Y., Lee, H. M., Ho, J. M. (2007). *Extracting citation relationships from web documents for author disambiguation*. Technical Report No. TR-IIS-06-017. Retrieved Oct. 4, 2010, from Retrieved Nov. 27, 2009, from http://www.iis.sinica.edu.tw/page/library/TechReport/tr2006/tr06017.pdf

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 Correlation. In *Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries. Lecture Notes In Computer Science*, (5173), p185 – 196. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://www.iis.sinica.edu.tw/papers/hoho/7642-F.pdf

Table 1: 28 Feature Combinations

	7 Combinations	21 Combinations with Features Y and P
Single-feature	C; T; J	CY; CP; CYP; TY; TP; TYP; JY; JP; JYP
Double-feature	CT; TJ; CJ	CTY; CTJ; CTP; TJY; TJP; TJYP; CJY; CJP; CJYP
Triple-feature	CTJ	CTJY; CTJP; CTJYP

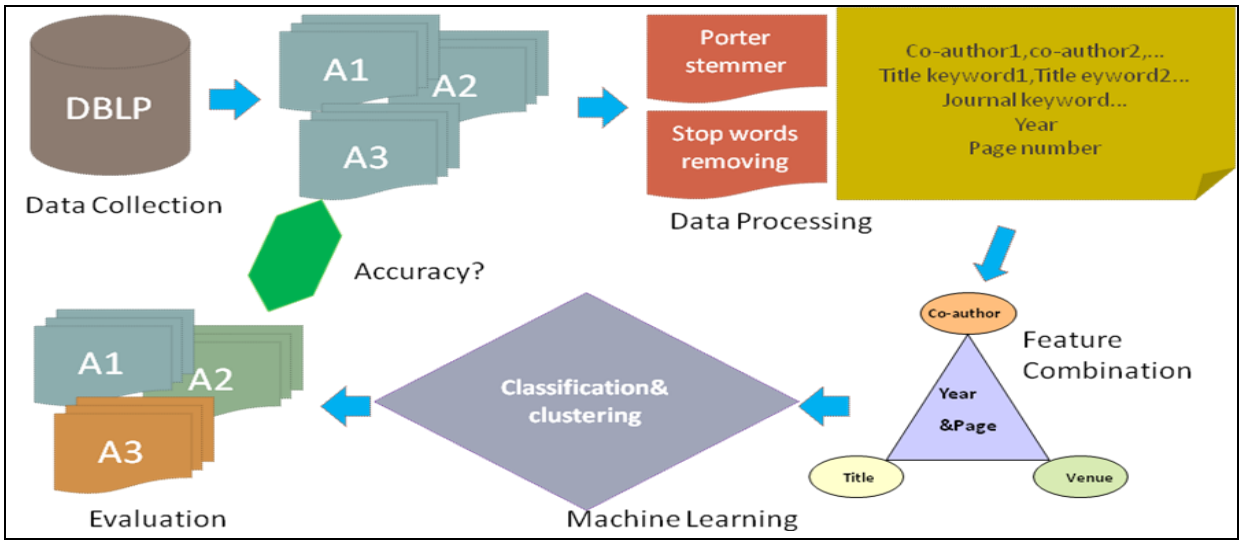

Figure 1: Research Procedure

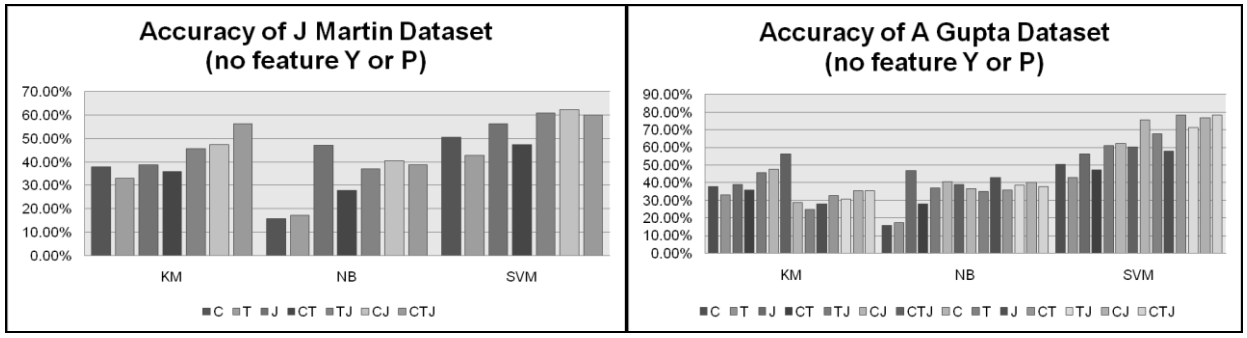

Figure 2 & 3: Performances of J. Martin Dataset and A. Gupta Dataset without features Y and P.

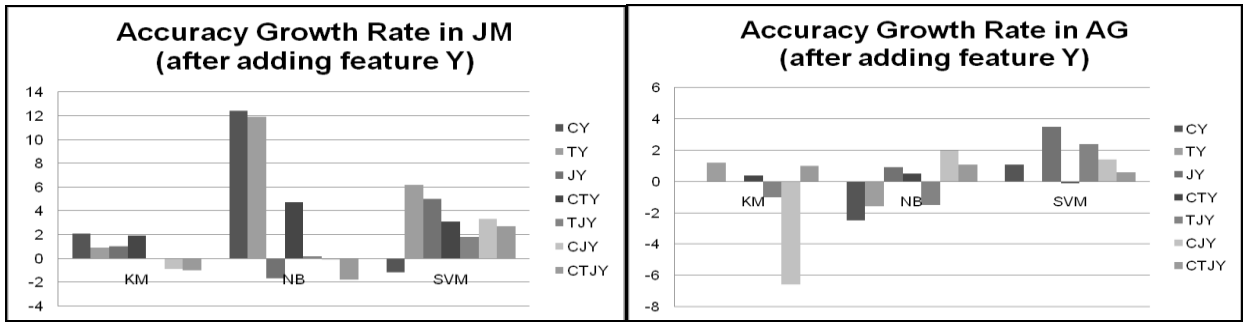

Figure 4 & 5: Performances of J. Martin Dataset and A. Gupta Dataset with features Y.

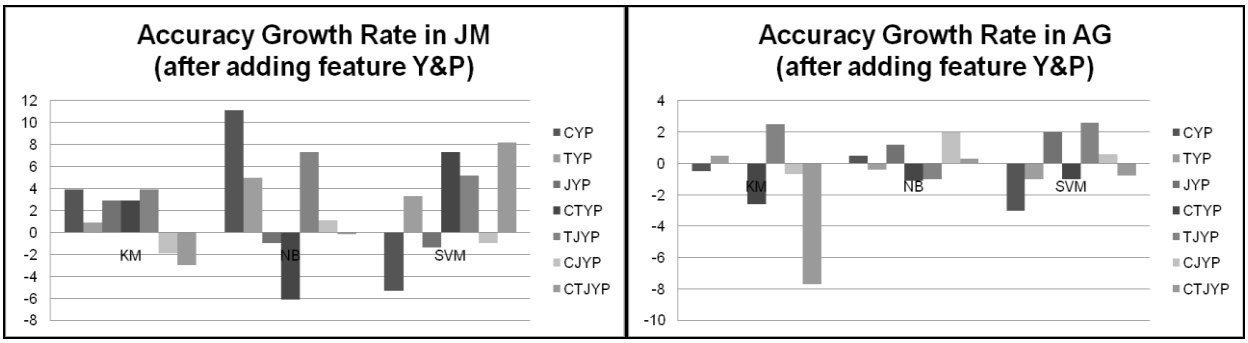

Figure 6 & 7: Performances of J. Martin Dataset and A. Gupta Dataset with features Y and P.