教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 49 , no. 2 (Winter 2011) : 215-240

Ambiguity Resolution for Author Names of Bibliographic Data

Kuang-Ha Chen
Professor
E-mail: khchen@ntu.edu.tw

Chi-Nan Hsieh
Graduate Student
E-mail: r97126004@ntu.edu.tw

Journal of Educational Media & Library Sciences 49 : 2 (Winter 2011) : 215-240

Research Article

Ambiguity Resolution for Author Names of Bibliographic Data[ψ]

Kuang-Hua Chen*

Professor
E-mail: khchen@ntu.edu.tw

Chi-Nan Hsieh

Graduate Student
Department of Library and Information Science
National Taiwan University
Taipei, Taiwan
E-mail: r97126004@ntu.edu.tw

Abstract

Users have been confronted with serious problems in ambiguities of author names, while a great deal of scholar information quickly accumulated in Internet. Therefore researches on ambiguity resolution for author name are indispensable. With comparison to previous work, this study attempts to address the problem using information contained in bibliographic data only. Five features, co-author (C), article title (T), journal title (J), year (Y), and number of pages (P), are used in this study. Note that feature Y and feature P are not ever used before. Both supervised learning methods (Naïve Bayes and Support Vector Machine) and unsupervised learning method (K-means) are employed to explore 28 different feature combinations. The findings show that the performance of feature journal title (J) and co-author (C) is very effective. Feature J plays an important role in three different methods, and feature C is effective in SVM. In addition, feature Y and feature P obviously enhance accuracy and the average improvement rate of feature Y is more significant than that of feature P (+2.5% in average). It is also shown that the performance of feature combination CTJ is not superior to JYP, and the performance of feature combinations CJY, JY and J are also very effective in the three methods. Finally, it is found that the accuracy of disambiguation on larger datasets is 10% inferior to that of the smaller ones, which indicated the limitation of using bibliographic data only. Consequently, the effective approach to disambiguating author name has to not only fully use bibliographic data but also introduce appropriate outer resources.

Keywords: *Author disambiguation, Bibliographic data, Machine learning*

Introduction

In general, names seem helpful in identifying a person with great ease. However, with widespread use of digital information in Internet era, name

---

ψ Part of this article had been presented at The International Conference of the 40th Anniversary of *Journal of Educational Media & Library Sciences*, March 7-8, 2011.

* Principal author for all correspondence.

2011/8/16 received; 2012/02/09 revised; 2012/02/09 accepted

ambiguity problems have commonly occurred. The ambiguity occurs in names with their abbreviated forms, typos, misspellings, multiple authors sharing the same name, or 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. Han et al. (2004) points out two types of common name ambiguities. The first type of name ambiguity occurs when an author has multiple name labels. For example, the author "David S. Johnson" may appear in various publications using different name abbreviations, such as "David Johnson", "D. Johnson", or "D. S. Johnson". The second one is that several authors may share the same name label. For instance, "D. Johnson" may refer to "David B. Johnson" from Rice University, "David S. Johnson" from AT&T research lab, or "David E. Johnson" from Utah University.

Many authorities are making their ways towards the problem. International Standard Organization has established International Standard Name Identifier (ISNI, 2010) and the Draft ISO Standard (ISO 27729) has planned to identify every creator of works by using unique 16-digital number. In addition, there are more and more nation-level systems developed in preparation for the coming of ISNI, such as Digital Author Identifier (DAI, 2010) in the Netherlands, People Australia (2010) service by the national library of Australia, and Research Name Resolver (2010) in Japan. Although the standard will take effect in the near future, lots of bibliographic documents and information with name ambiguities still need to be coped with.

In fact, many well-known database vendors also contribute to solutions to the pressing problem. Two approaches are usually applied to handling this problem. The first approach is to build supplementary identification functionalities to help end-users to identify their retrieval results. Elsevier (2010), for instance, provides "author search" function for its Scopus Database. The function can help users search ambiguous names and make a list of these authors sharing the same name label. However, it still requires complete author information to produce desired results, such as affiliation, subject area, or resident city/country of these authors. Besides, Web of Science database by Thomson Reuters (2010) offers Distinct Author Identification System, which claims it uses proprietary algorithm to cluster the namesakes and his/her works. However, the system does not process every record in database (only before 2007), and the performances of its clustering is unknown. The second one is to establish a registry of unique author identifiers, such as Researcher ID by Thomson Reuters (2010) and Author Service by Wiley-Blackwell (2010). Even if the mechanism looks simple and feasible, they are in fact passive methods. Different identifiers may still make users feel more confused.

Libraries usually build or apply authority files in response to these ambiguities, such as OCLC (2010) WorldCat Identity Service and the Scholar Universe of ProQuest (2010). The former service contains more than 20 million name records, but it is just in its beta version so far. The latter also provides high-quality name search by the professional editor group of ProQuest, and it offers two millions profiles to users for free. 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.

In general, the background mentioned above shows that name or author disambiguation is not complicated when it comes with sufficient and correct individual information. In reality, however, the personal information is not easily available. Therefore, this study attempts to identify authors sharing same name by using bibliographic data only, which is generally available in bibliographic databases or digital libraries. Two objectives of this study are: 1) to explore how the performance can be achieved by using bibliographic data only, which is composed of authors, article titles, journal titles, publication date and number of pages and 2) to investigate the effectiveness of features publication date and number of pages, which have never been discussed before.

The structure of this paper is shown as follows. Section 2 describes previous studies. Section 3 introduces methodology of this study. Section 4 presents the experimental results and discusses the findings. Section 5, finally, gives a brief conclusion.

Previous Work

This study focuses on ambiguity resolution for author in bibliographic data. Name disambiguation, in general, will be discussed first in this section. After general discussion to name disambiguation, disambiguation for author name will be discussed to have a fundamental understanding on this issue.

Name disambiguation

The problem of name ambiguity originates in a broader issue: identity uncertainty and the study of pioneers in this area called “record linkage” by Fellegi and Sunter (1969). They developed a statistical model to process multiple records in databases and regard records as feature vectors in order to measure their similarity. This approach has influences on several studies related to database managements, such as data merge/purge (Hernandez & Stolfo, 1998) and duplicate record detection (Elmagarmid et. al., 2007). Nowadays, digital library researchers and large-scale database vendors have not only paid attention to keywords search but also emphasized the importance of name/author search (Smalheiser & Torvik, 2009). Therefore, name disambiguation has been received much more attention in recent years.

In general, to carry out name disambiguation, just like data or text mining, a "machine learning" model has to be constructed (Mitchell, 1997). Machine learning depends on the "training set" to select important features and then the trained model is used to determine the class of target items. Finally, appropriate methods of evaluation will be carried out, which would be discussed further later. Two sorts of machine learning approaches are considered in name disambiguation: supervised and unsupervised learning. The key difference between supervised methods and unsupervised methods is that supervised learning methods need labeled data for training, while unsupervised methods do not. The performance of supervised methods is generally better than that of unsupervised one. In the work of disambiguating authorship, each author name can be considered as a class and then name disambiguation classifies citations into their author classes (Han et al., 2005a).

Many researchers have developed related mechanisms or procedures for name disambiguation in recent years, but the datasets they used are not identical. The diversities of datasets influence the types of selected features and the methods for evaluation. More features considered, in general, could have higher possibility to achieve better performance, so the researchers presently look for new sources of features. However, there are still many alternatives to resolutions of name ambiguity using the same features. Some put emphasis on the distance between strings (Torvik et al., 2005), and others emphasized the use of prior knowledge (French, Powell, & Schulman, 2000). Moreover, different methods for feature weighting are proposed in literature, such as Jaccard, TFIDF (Term Frequency and Inverse Document Frequency), Jaro-Winkler and Levenstein, and so on.

Several studies show the current status of name disambiguation. Authorship attribution and stylometry via the signatures of writing have applied to the study about the novelist's change of literary style over time (Can & Patton, 2004) and prediction of an author's gender (Koppel et al., 2002). Record linkage in administrative databases has a long history based on the work by Fellegi and Sunter (1969). A number of follow-up researches are constantly implemented for various data, such as public health records (Jaro, 1995), census records (Winkler, 1995), name and address information (Churches et. al., 2002). Ambiguity resolution for authors has developed in recent years. Several research groups used different sources of dataset, such as bibliographic data (e.g. Hill & Provost, 2003; Han et al., 2004, 2005a, 2005b; Huang, Ertekin, & Giles, 2006; Bhattacharya & Getoor, 2007; Culotta et. al., 2007), the parts of full-texts (Song et al., 2007), and the information of web pages (e.g. Kanani et al., 2007; Yang et al, 2007, 2008; Tan, Kan & Lee, 2006). The applications to the records in multimedia database are active as well, such as automatically building authority file of sheet music (DiLauro et al., 2001) and name disambiguation for Internet Movie DataBase (IMDB) by social network model of individuals (Malin, Airoldi & Carley, 2005).

Ambiguity resolution for author names has been the focus of general name disambiguation in many realistic researches. Therefore, we will discuss author name disambiguation in detail in the next subsection.

Ambiguity resolution for author

As mentioned above, several research task forces devoted themselves in author name disambiguation for different purposes. "CiteSeer" is a famous digital library service developed by Steve Lawrence, Lee Giles and Kurt Bollacker (CiteSeer, 2011). CiteSeer collected documents to establish a full-text database using web crawlers. Maintaining correctness and consistence of data in a large-scale database demands appropriate algorithms and automatic classification or clustering.

Earlier studies stressed the methods of classification/clustering and computerized scalability by using limited feature combination (i.e. co-author, title and journal title), so accuracy was not the first concern (Han et al., 2004, 2005a, 2005b; Huang, Ertekin, & Giles, 2006). Later studies managed to apply additional features, such as the first page of the paper.

Getoor and his colleagues (2006, 2007), then, emphasized the analysis of author social network. In the beginning, Bhattacharya and Getoor (2006) used LDA to cluster bibliographic records based on name tokens, but the implementation process is too time-consuming. They introduced in the concept of "collective entity resolution" and found that recognition results can help each other. For example, assume name *A* and name *B* co-occurred in two records. If it has been confirmed that two *A*s are different individuals, it is probable to infer that two *B*s are also different persons (Bhattacharya & Getoor, 2007). In contrast, Bilgic et al. (2006) developed an interactive disambiguation system "D-Dupe", which used bibliographic information to build a co-authorship network in order to assist in the manual identification.

McCallum and his colleagues have published a series of influential studies in author disambiguation and created a digital library called Rexa, which contains seven million records of computer science literature. The characteristics of their works are three-way and high-order simultaneous comparisons (beyond common pairwise comparisons). Culotta et al. (2007) employed aggregate constraints to enhance their model based on article titles, emails, affiliations and venue of publication, etc. Kanani, McCallum, and Pal (2007) exploited active learning for web information gathering in order to supplement articles' metadata. That is to say, applying any available resource for author name disambiguation is one of mainstreams in this research field.

In Han's studies (Han et al., 2004, 2005a, 2005b), they first constructed a test suite (hereafter DBLP dataset) using bibliographic records of DBLP database.

Supervised methods and unsupervised methods were then used for author name disambiguation. The former achieved accuracy of 70%, and the latter 65%. However, only co-author names, article titles, and journal titles were used in their study. Yang et al. (2007, 2008) subsequently used the same dataset by Han et al. (2005a) and added outside features from web to their disambiguation work by pair-wise clustering. Yang et al. (2007) extracted citation relationships from the URL information of web document, and they improved the method by building topic and web correlation (Yang et al., 2008). Eventually, the accuracy of Yang's results (2007, 2008) is better than Han's in general. Table 1 shows the comparisons of their performance. However, the web information on the Internet is not always available and requires additional manual work.

Table 1 Summary of Previous Work

Researcher	Method	Dataset	Accuracy
Han et al. (2004)	Two Supervised Learning Approaches (Bayes vS. SVM)	1) Publication in author homepages (2 names) 2) Citation in DBLP database (9 names)	1) 94.5% (SVM better) 2) 73.3% (Bayes better)
Han et al. (2005a)	Hierarchical Naïve Bayes mixture model	1) Publication in author homepages (2 names) 2) Citation in DBLP database (14 names)	1) 65.5% 2) 63.2%
Han et al. (2005b)	K-way Spectral Clustering	1) Publication in author homepages (2 names) 2) Citation in DBLP database (14 names)	1) 71.2%, 84.3% 2) 61.5%-64.7%
Yang et al. (2007)	Pair-wise clustering with additional web information	Citation in DBLP database (14 names)	20% better than Han's K-way
Yang et al. (2008)	Pair-wise clustering with additional topic & web correlation	Citation in DBLP database (14 names)	25% better than Han's K-way

In general, each method or approach mentioned above could be applied to any database with bibliographic data, such as DBLP, CiteSeer, arXiv, MEDLINE, Google Scholar, Web of Science (Thomson Scientific), Scopus (Elsevier), ADS (Astrophysics Data System), Libra (Academic Search), and RePEc. In addition to bibliographic data, some outer resources are taken into account for delivering satisfactory performance as well, such as full-text articles and information from web pages. However, copyright of full-texts and privacy concerns of author information could be a hindrance to obtaining these supplementary resources. For these reasons, we consider author name disambiguation using information contained in bibliographic data only and would like to investigate the feasibility and performance based on this consideration accordingly.

Therefore, the purpose of this study is to explore performance of various feature combinations using complete information of bibliographic data and

investigate influences of features which were not used ever before, i.e., “year” and “number of pages”, on disambiguation.

Research Design

In order to investigate different factors, e.g., feature combinations, learning methods, and complexities of datasets, many resources are used and arranged in this study. The research framework is shown in Figure 1. The procedure consists of data collecting, data processing, model learning, and performance evaluating. The following subsections explain these stages.

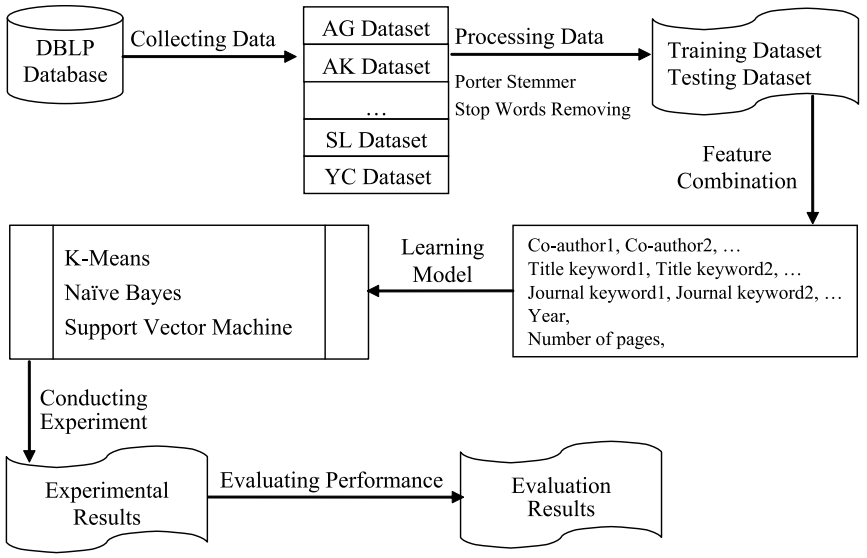


Figure 1 Research Procedure

Collecting data

The datasets employed in this study was the same DBLP datasets constructed by Han et al. (2005a, 2005b), which contains 8,441 bibliographic records collected from DBLP database. The datasets consists of 14 popular author names shared by 476 individual authors. In order to increase the complexity of ambiguity, the first names of author names were changed into initials in Han’s design. The DBLP datasets of this study is provided by Dr. Giles, but the feature information that we would like to analyze consists of five features (i.e. co-authors, article titles, journal titles, year and number of pages) rather than three features which Han et al. (2005a, 2005b) used in their study.

Therefore, we have to supplement the needed features, i.e., year and number of pages. In the process of data supplementing, we unfortunately found some problems of the DBLP datasets as the failure cases pointed by Pereira et al. (2009), such as wrong author names or duplicate names marked in bibliographic record,

the lack of article titles or journal titles. We then have to revise and delete some bibliographic records in DBLP datasets accordingly. The statistics of test data used in this study is shown in Table 2.

Table 2 The Five Ambiguous Author Name Datasets

Name	Number of Different Authors		Number of Bibliographic Records	
	Original	**Revised**	Original	**Revised**
A. Gupta (AG)	26	**26**	577	**572**
A. Kumar (AK)	14	**14**	244	**238**
C. Chen (CC)	61	**61**	800	**679**
D. Johnson (DJ)	15	**15**	368	**347**
J. Lee (JL)	100	**99**	1417	**1270**
J. Martin (JM)	16	**15**	112	**103**
J. Robinson (JR)	12	**12**	171	**168**
J. Smith (JS)	30	**29**	927	**872**
K. Tanaka (KT)	10	**10**	280	**267**
M. Brown (MB)	13	**13**	153	**146**
M. Jones (MJ)	13	**13**	259	**247**
M. Miller (MM)	12	**12**	412	**384**
S. Lee (SL)	83	**84**	1457	**1260**
Y. Chen (YC)	71	**71**	1294	**1168**
Total	476	**474**	8471	**7720**

Processing data and feature combinations

The purpose of this study focuses on performance of complete combinations of various features (e.g. authors, article titles, journal titles, venues) in bibliographic data for disambiguation. Accordingly 28 feature combinations are explored in the study to examine how each 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 "never-used" features Year (Y) and Number of Pages (P). The possible combinations are shown in Table 3.

Table 3 28 Feature Combinations

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

Of course, a few pre-processing tasks are considered in our study. Porter's stemmer is used for titles (feature T) and journal titles (feature J), and stop words are removed by stop-words corpus from Toolkit in NLTK. In this way, it is believed that the remaining words in those two features are meaningful keywords.

Besides, the word occurrence is also considered for feature weighting which has been considered by many information retrieval researches (Lu, Xu, & Geva, 2008), so TFIDF scheme is adopted in the work of data processing. Term Frequency (TF) stands for the frequency of occurrence of keyword term in the

bibliographic record, and Inverse Document Frequency (IDF) stands for the inverse of the frequency of occurrence of keyword term in the dataset.

Learning model and evaluating performance

After data processing, each bibliographic record is transformed into a feature vector and ready for classification or clustering. Both supervised learning methods and unsupervised learning methods are employed to examine the performance of author name disambiguation. Two supervised learning methods used are Naïve Bayes (Toolkit in NLTK) and Support Vector Machine (LIBSVM) (Chang & Lin, 2010). The input format of Naïve Bayes in NLTK is "index = value". In addition, the format of SVM by LIBSVM is "index: value", and the attribute with null value in records is deleted. Both tools automatically generate accuracy value for evaluation. The ratio of training set and testing set is 7 : 3 and cross validation is used in training process.

For unsupervised learning method, K-means clustering is conducted using cluster module of Python. The input format of the K-means cluster module is vector tuple, such as "(5, 3), (10, 3)". Besides, the number of clusters is based on heuristics of our pretest implementation. Two author name datasets, A. Gupta and C. Chen, are used in pretest. We gradually increase the number of clusters from 5 to 150. Finally, we find while the number of authors of the dataset is fewer than 60, we will run K-means clustering from 5 clusters to 60 clusters. If the number is more than or equal to 60, we will run from 60 to 125. After clustering, the decision of label of each cluster is based on the number of tuple in cluster.

Like Han et al. (2005a, 2005b) and Yang et al. (2007, 2008), we evaluate the performance in terms of the disambiguation accuracy, calculated by dividing the sum of correctly clustered bibliographic records by the total number of bibliographic records in the dataset. The disambiguation accuracy is defined as follows:

$$Accuracy = \frac{\sum_{i \in I} n_{ir}}{N}$$

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 bibliographic records in the dataset.

Settings for year and number of pages

In order to consider features Year (Y) and Number of pages (P) in the study, year and number of pages in bibliographic data have to be transformed into corresponding codes meaningfully. For feature Year (Y), it is assumed that each author has his/her period of academic production, so year distribution of the whole dataset is segmented into intervals. According to the dataset, the publication dates of literature in DBLP were mainly between 1975 and 2005. Based on this observation, a time span of 10 years is used in this study.

As for number of pages (P), under the influence of publication types and authors' preference, numbers of pages of the bibliographic data are calculated first and intervals are set based on number of pages conventions of different types of publications. For example, the average length of papers of top 15 journals of computer science in Journal Citation Report (Thomason Routers, 2011) is 16.6 (see Table 4).

Table 4 The Length of Regular Paper in Top 15 CS Journals (up to Jan 2011)

Rank	Abbreviated journal title	Length of paper	5-year impact factor
1	*ACM COMPUT SURV*	35	**7.667**
2	*HUM-COMPUT INTERACT*	8	**6.190**
3	*COMPUT INTELL*	12	**5.378**
4	*IEEE T EVOLUT COMPUT*	No proclaimed specially	**4.589**
5	*VLDB J*	25	**4.517**
6	*MIS QUART*	20	**4.485**
7	*IEEE T PATTERN ANAL*	14	**4.378**
8	*J AM MED INFORM ASSN*	10	**3.974**
9	*J CHEM INF MODEL*	No proclaimed specially	**3.882**
10	*J COMPUT AID MOL DES*	No proclaimed specially	**3.835**
11	*IEEE T SOFTWARE ENG*	14	**3.750**
12	*ACM T GRAPHIC*	No proclaimed specially	**3.619**
13	*IEEE T MED IMAGING*	8	**3.540**
14	*INT J COMPUT VISION*	No proclaimed specially	**3.508**
15	*J WEB SEMANT*	20 (from 15 to 25)	**3.412**
		Average = 16.6 =>17	

Three segmented points are designed in the study: three pages for poster papers, eight pages for conference papers, and more than 17 pages for journal papers. Then four intervals are 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 number of pages were considered.

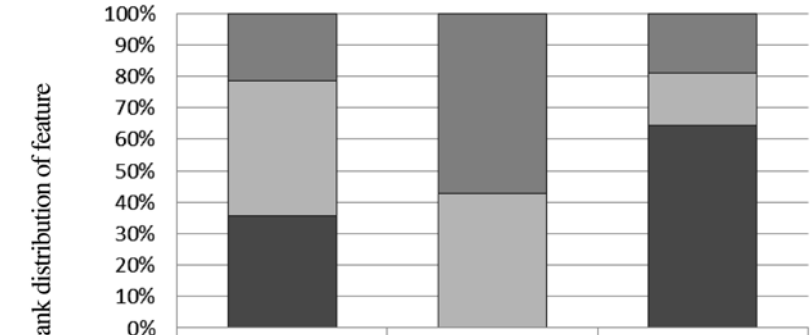
Experimental Results

In this study, 14 author names of DBLP datasets are examined (see Table 2 above). Each feature combination is investigated with particular focus on features Y and P. In addition, the complexity of datasets is also explored. In the end, the features (or feature combinations) achieving best performance in each dataset are highlighted.

Common feature combinations

To begin with, the performance of author disambiguation without considering features Y and P is described. Because of the following comparisons of various feature combinations are considered three methods in this study, the statistics of rank are based on comparisons of 42 times (combinations of 14 datasets and three methods).

In one-feature (C, T and J) experiment, feature J scored 64.2% of the lead in the experiments (see Figure 2). Feature C obtained 37.5% of the lead, but feature T did not obtain the lead ever. This indicates that the outstanding performance of feature J and feature C in the disambiguation work for authors, and feature J is only satisfactory. In two-feature (CT, TJ and CJ) experiment, feature CJ scored 78.5% of the lead in the experiments (see Figure 3). Then, feature TJ obtained 19.0% of the lead, but feature CT only achieved 7.1% of the lead. As the result of comparison in one-feature (J > C > T), the rank comparison of two-feature is not surprising (CJ > TJ > CT).



	C	T	J
The Worst	9	24	8
The Runner-up	18	18	7
The Lead	15	0	27

Figure 2 Rank Comparisons of Single Feature

However, it is found that the rank comparison of each feature combination is to a large extent influenced by different methods. Please take a look at the rank of one-feature in Table 5. Feature J achieves the first rank in K-means clustering (KM for short) and Naïve Bayes (NB for short) steadily. In contrast, the performance of feature C is generally more desired than feature J in Support Vector Machine (SVM for short). Then, in the rank of two-feature, although feature CT is always the worst in KM and NB, it is also not the case in SVM.

In three-feature (CTJ) experiment, it is concerned that whether CTJ achieved the best accuracy in the dataset owing to CTJ commonly regarded as "default" feature combination in many previous works. It is shown that feature CTJ leads other feature combinations only 7 times in the 42 times of comparisons of the best accuracy, and 6 times out of 14 times in SVM. As a result, when features C, T, and J are used at the same time, the combination cannot necessarily ensure the best performance. The performance of feature combination CTJ in SVM is different from KM and NB. In fact, the results in SVM match the findings of the study by Han et al. (2004).

Table 5 Statistics of Rank Comparisons in Different Methods

K-means (KM)

Rank of Single-Feature				Rank of Two-Feature				Best Accuracy	
	C	T	J		CT	TJ	CJ		CTJ
A. Gupta	2	3	1	A. Gupta	3	1	2	A. Gupta	no
A. Kumar	2	3	1	A. Kumar	3	2	1	A. Kumar	no
C. Chen	3	2	1	C. Chen	3	2	1	C. Chen	no
D. Johnson	2	3	1	D. Johnson	3	1	2	D. Johnson	no
J. Lee	2	3	1	J. Lee	3	1	2	J. Lee	no
J. Martin	2	3	1	J. Martin	3	2	1	J. Martin	no
J. Robinson	1	3	2	J. Robinson	2	3	1	J. Robinson	no
J. Smith	2	3	1	J. Smith	3	2	1	J. Smith	no
K. Tanaka	3	2	1	K. Tanaka	3	1	2	K. Tanaka	yes
M. Brown	1	3	2	M. Brown	3	2	1	M. Brown	no
M. Jones	1	3	2	M. Jones	2	1	3	M. Jones	no
M. Miller	2	2	1	M. Miller	1	1	1	M. Miller	no
S. Lee	2	3	1	S. Lee	3	2	1	S. Lee	no
Y. Chen	2	3	1	Y. Chen	3	2	1	Y. Chen	no

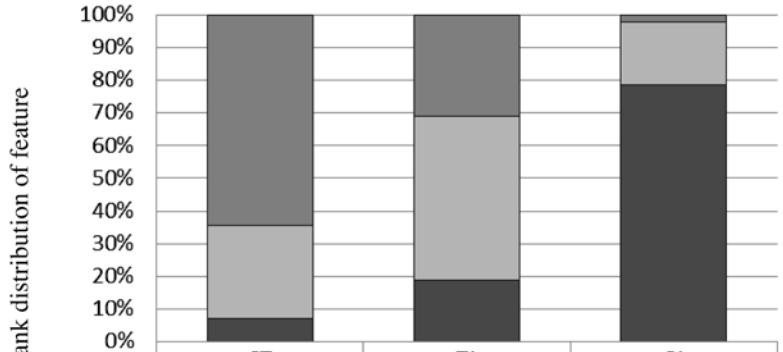
Naïve Bayes (NB)

Rank of Single-Feature				Rank of Two-Feature				Best Accuracy	
	C	T	J		CT	TJ	CJ		CTJ
A. Gupta	2	3	1	A. Gupta	3	2	1	A. Gupta	no
A. Kumar	3	2	1	A. Kumar	3	2	1	A. Kumar	no
C. Chen	2	3	1	C. Chen	3	2	1	C. Chen	no
D. Johnson	3	2	1	D. Johnson	3	1	2	D. Johnson	no
J. Lee	2	3	1	J. Lee	3	2	1	J. Lee	no
J. Martin	3	2	1	J. Martin	3	2	1	J. Martin	no
J. Robinson	2	3	1	J. Robinson	3	2	1	J. Robinson	no
J. Smith	2	3	1	J. Smith	3	2	1	J. Smith	no
K. Tanaka	2	3	1	K. Tanaka	3	2	1	K. Tanaka	no
M. Brown	1	3	2	M. Brown	2	3	1	M. Brown	no
M. Jones	3	2	1	M. Jones	3	2	1	M. Jones	no
M. Miller	1	3	2	M. Miller	2	3	1	M. Miller	no
S. Lee	2	3	1	S. Lee	3	2	1	S. Lee	no
Y. Chen	2	3	1	Y. Chen	3	2	1	Y. Chen	no

Support Vector Machine (SVM)

Rank of Single-Feature				Rank of Two-Feature				Best Accuracy	
	C	T	J		CT	TJ	CJ		CTJ
A. Gupta	1	2	3	A. Gupta	1	3	2	A. Gupta	yes
A. Kumar	3	2	1	A. Kumar	3	2	1	A. Kumar	no
C. Chen	1	2	3	C. Chen	2	3	1	C. Chen	no
D. Johnson	1	2	3	D. Johnson	2	3	1	D. Johnson	no
J. Lee	1	2	3	J. Lee	1	3	2	J. Lee	yes
J. Martin	2	3	1	J. Martin	3	2	1	J. Martin	no
J. Robinson	1	3	2	J. Robinson	2	3	1	J. Robinson	no
J. Smith	1	3	2	J. Smith	2	3	1	J. Smith	yes
K. Tanaka	1	2	3	K. Tanaka	3	2	1	K. Tanaka	yes
M. Brown	1	2	3	M. Brown	2	3	1	M. Brown	yes
M. Jones	3	2	1	M. Jones	3	1	2	M. Jones	yes
M. Miller	3	2	1	M. Miller	2	3	1	M. Miller	no
S. Lee	1	2	3	S. Lee	2	3	1	S. Lee	no
Y. Chen	1	2	3	Y. Chen	2	3	1	Y. Chen	no

Note: 1 = the lead, 2 = the runner-up, 3 = the third;
yes / no= Whether CTJ achieved the best accuracy in the dataset



	CT	TJ	CJ
The Worst	27	13	1
The Runner-up	12	21	8
The Lead	3	8	33

Figure 3 Rank Comparisons of Two Features

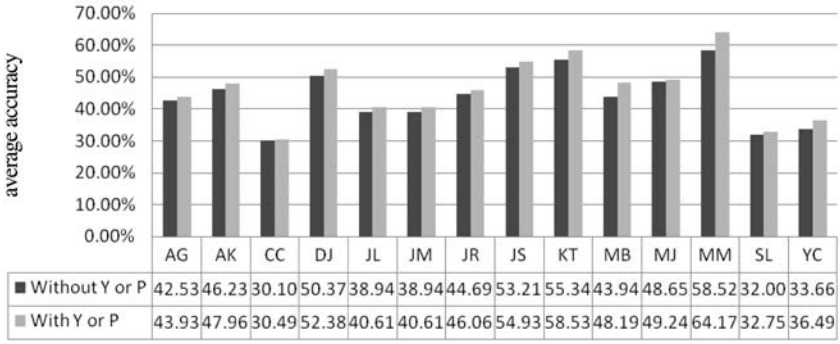
Features year (Y) and number of pages (P)

In order to present the influence of features Y and P, the average performance of each feature combination is shown in Figure 4. In contrast, the average improvement rates of performance with considering features Y, P or YP are investigated and shown in Figure 5. These results indicate that the performance with using features Y and P is better than that without using Y and P.

However, the performance above mentioned is estimated by the average accuracy rates in three methods. Therefore, separate performance with inclusion of feature Y and P is discussed as follow. The different impacts with inclusion of feature Y and feature P by three methods are shown in Figure 6 and Table 6. The improvement rate, which is the difference between the performance without and with feature Y or feature P, is examined in this section.

First, with the inclusion of feature Y, the average improvement rates in KM are 6.08% (sd = 6.76%), 0.73% (sd = 1.00%) in NB model and 0.49% (sd = 1.12%) in SVM, respectively. Then, after adding feature P for author name disambiguation, the average improvement rates in KM are 3.59% (sd = 4.09%), 0.59% (sd = 0.82%) in NB model and –0.39% (sd = 0.95%) in SVM. Finally, when features Y and P are included at the same time, the average improvement rates in KM are 5.21% (sd = 5.28%), 1.38% (sd = 1.67%) in NB model and 0.33% (sd = 0.98%) in SVM (see Table 6).

From the findings shown above, feature Y and feature YP obviously delivered positive performance in our datasets. In addition, the inclusion of feature P also produced positive effects, but the influence is not obvious. However, the effect is more positive in K-means clustering (+4.98% in average) than that in Naïve Bayes Model (+0.90% in average) and Support Vector Machine



	AG	AK	CC	DJ	JL	JM	JR	JS	KT	MB	MJ	MM	SL	YC
Without Y or P	42.53	46.23	30.10	50.37	38.94	38.94	44.69	53.21	55.34	43.94	48.65	58.52	32.00	33.66
With Y or P	43.93	47.96	30.49	52.38	40.61	40.61	46.06	54.93	58.53	48.19	49.24	64.17	32.75	36.49

Figure 4 The Comparison Using with/out Features Y and P
(Average in 3 Methods)

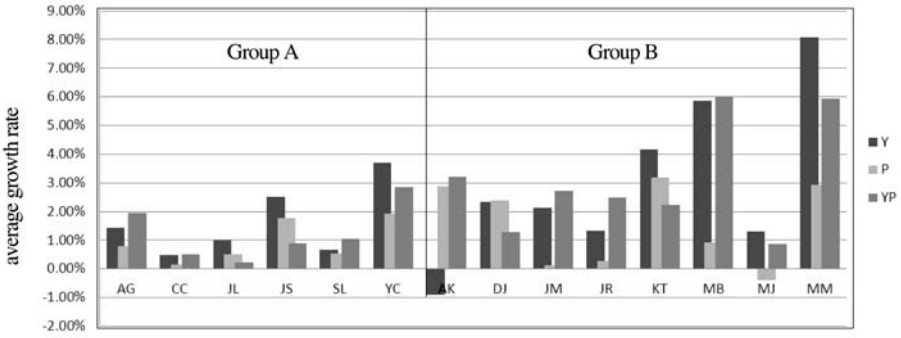


Figure 5 Average Improvement Rate Using Features Y and P
(Average in 3 Methods)

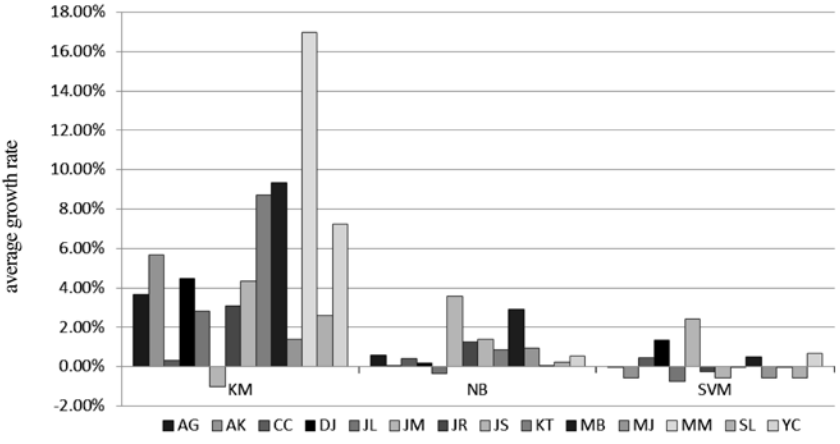


Figure 6 Improvement Rate Using Features Y and P
(Average of Y, P and YP)

Table 6 Improvement Accuracy Rate with the Inclusion of Feature Y and P

	KM			NB			SVM		
	Y	P	YP	Y	P	YP	Y	P	YP
AG	2.89	3.16	4.99	0.47	0.63	0.60	0.97	–1.43	0.30
AK	–1.24	9.53	8.81	0.07	–0.13	0.17	–1.57	–0.77	0.69
CC	0.43	0.41	0.13	0.10	–0.11	1.19	0.89	0.17	0.24
DJ	5.69	5.69	1.19	0.11	0.01	0.41	1.21	0.59	2.27
JL	3.20	3.16	2.07	–0.27	–0.63	–0.09	0.06	–1.03	–1.29
JM	0.86	–3.73	–0.13	2.70	1.91	6.10	2.87	2.21	2.20
JR	2.97	1.53	4.77	0.86	0.66	2.29	0.19	–1.36	0.43
JS	6.44	5.51	1.09	1.50	0.79	1.91	–0.40	–1.03	–0.31
KT	10.14	9.64	6.33	1.41	0.69	0.56	0.93	–0.77	–0.23
MB	13.64	0.23	14.19	2.67	2.54	3.46	1.24	–0.01	0.29
MJ	3.94	–1.56	1.84	0.56	0.89	1.34	–0.57	–0.54	–0.61
MM	24.79	8.59	17.50	–0.53	0.24	0.36	–0.06	0.00	–0.03
SL	2.23	2.37	3.19	0.23	0.20	0.24	–0.46	–0.99	–0.26
YC	9.16	5.70	6.91	0.37	0.50	0.80	1.61	–0.43	0.86
Avg.	6.08	3.59	5.21	0.73	0.59	1.38	0.49	–0.39	0.33

(+0.15% in average). Please refer to Figure 6. It is shown that feature Y and feature P could significantly enhance performance in K-means clustering, but not obviously in Naïve Bayes and SVM. In the experiment of K-means clustering, the improvement rate with feature Y maximally achieve 24.79% in MM Dataset, and feature P achieve 9.53% in AK Dataset and feature YP achieve 17.5% also in MM Dataset. But the maximum of improvement with feature Y or P in the experiment of Naïve Bayes and Support Vector Machine is about 2.5% at most. It seems feasible to explore whether the feature Y and P could efficiently enhance accuracy in various unsupervised approaches in future studies.

Complexity of datasets

According to the scale of datasets, the datasets are separated into two groups: Group A and Group B. Group A contains complicated datasets (more than 20 individuals and more than 400 bibliographic records), such as A. Gupta, C. Chen, J. Lee, J. Smith, S. Lee and Y. Chen. Group B includes the less complicated datasets (fewer than 20 individuals and fewer than 400 bibliographic records), such as A. Kumar, D. Johnson, J. Martin, J. Robinson, K. Tanaka, M. Brown, M. Jones and M. Miller.

In fact, the performance of Group A is not as good as Group B. The average performance of Group A is 39.14%, but 49.62% in Group B. Moreover, it is obvious that the impact with feature Y and P in Group A is less significant than Group B. The average improvement rate of Group A is 1.28, but 2.56% in Group B. Please refer to Figure 5. These suggest that the complexity of datasets can influence the performance. In other words, ambiguity in much larger datasets increases quickly like the complexity in the real world.

Top one feature combinations

Feature combinations achieving the best accuracy are explored here. Table 7 shows the "top 1 feature combination" for different methods and different author name datasets. Figure 7 displays top 1 distribution for different feature combinations. As shown in Table 7 and Figure 7 below, the significance of feature JYP and CTJ is obvious. Note that J, JY and CJY are of the third, fourth and fifth place, respectively.

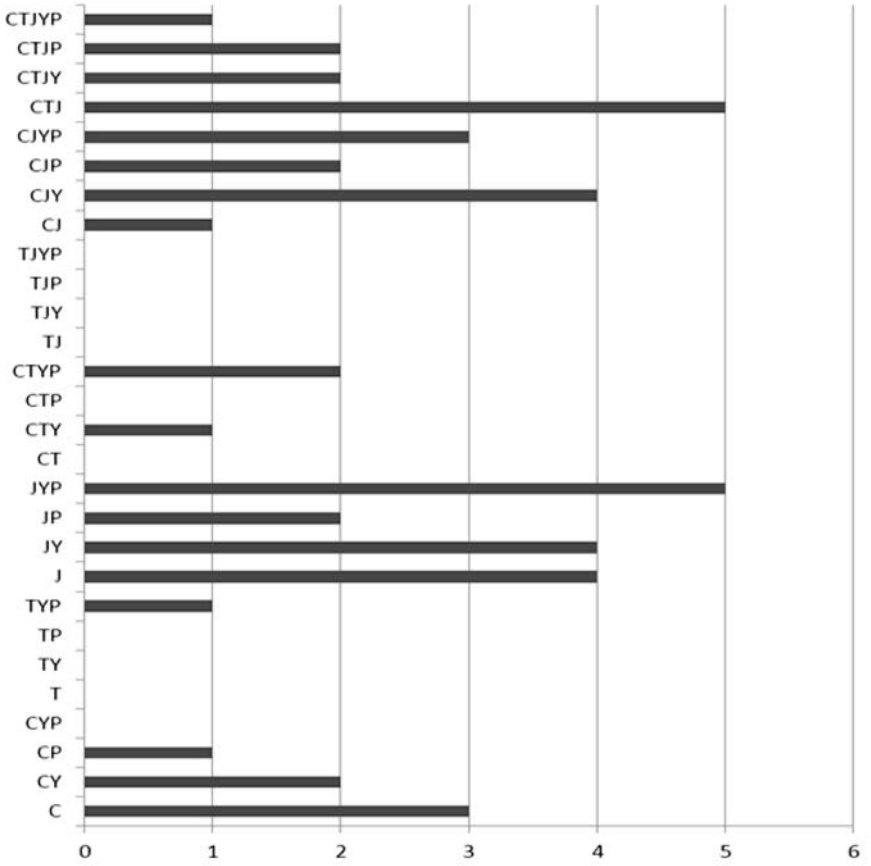


Figure 7 Top 1 Distribution of Feature Combinations

Table 7 shows 14 out of 18 top 1 feature combinations contain feature Y or feature P. That means features Y and P have their roles in author name disambiguation even though they have been not ever considered before. In addition, while considering distribution of each feature in all top 1 feature combinations, it is found that Y and P are not the worst. Please refer to Figure 8. Feature J accounted for 77.7% of top 1 feature combinations, feature C for 64.4% secondly, and feature Y thirdly.

Table 7 Top 1 Feature Combinations

	KM	NB	SVM
AG	CTJY	JY	CTJ
AK	CP	JY	CJYP
CC	J	JYP	CJY
DJ	JP	JYP	CTYP
JL	J	JP	CTJ
JM	J	JY	CJP
JR	C	JYP	CTJY
JS	CY	CJY	CTJ
KT	CTY	CJP	CTJ
MB	TYP, CTYP, TJYP, CJYP, CTJYP	C	CTJ
MJ	C	CJYP	CTJP
MM	JY	CJY	CTJP
SL	J	JYP	CJ
YC	CY	JYP	CJY

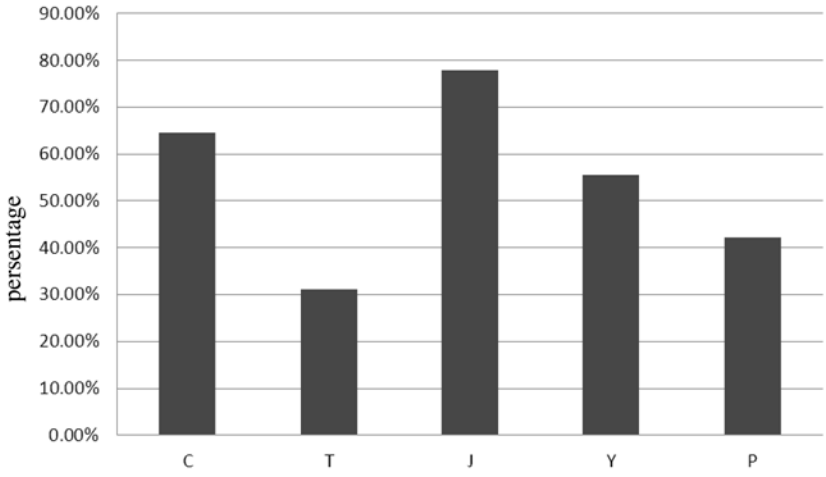


Figure 8 Percentage of Features in Top 1 Feature Combinations

Conclusions

This study investigates the effectiveness of features Y and P, and the performances of feature combinations on author disambiguation. It is shown that CTJ cannot necessarily ensure the best performance. In previous works, this common feature combination was usually regarded as a normal scheme, and these works focused on the designs of algorithm or the impacts of new resources. It is few to pay much attention to fully apply possible bibliographic features to author disambiguation. This study shows that the performance of JYP is not inferior to that of CTJ, and the performances of CJY, JY and J are also good in general. Although the best feature combination is mainly contributed by C and J, the inclusion of Y and P can substantially enhance the performance as well.

The inclusion of Y and P exhibits positive influence on disambiguation. The average improvement rates of the inclusion of Y, P, and YP are 2.44%, 1.29%, and 2.30%, respectively. As Section 4.2 mentioned, the impacts of Y and P are significant in K-means (improvement of accuracy is about 5%). However, the influence of them is not obvious in Naïve Bayes and Support Vector Machine. It seems feasible to explore if Y and P could efficiently enhance accuracy in various "unsupervised" approaches in future. In addition, the setting for Y and P ought to depend on characteristics of datasets. For example, the setting for number of pages for journals of humanities or social science should be more than 17.

Various feature combinations have different effects on author name disambiguation while using different clustering or learning methods. It is found that the performances of J and JYP in K-means clustering and Naïve Bayes Model are comparable to those of C and CTJ in SVM. Moreover, as the previous findings suggested, average improvement rate of Y and P in K-means (4.98%) is significantly better than that in Naïve Bayes (0.90%), but the improvement rate in SVM is not effective (0.15%). In other words, the "collocation" of features and learning approaches is an important research issue in author disambiguation.

The scale of datasets probably takes effects due to the different complexity. In general, the performances on larger datasets are much lower than those of the smaller ones, and the effectiveness is not obvious while introducing features Y and P. This reveals limitations of the solution of using bibliographic data only. As a consequence, using of appropriate outer resources is a critical issue for name or author disambiguation in future.

Acknowledgments

Parts of this work are supported by National Science Council under the grant number NSC98-2628-H-002-003-MY2.

References

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-36.

Can, F., & Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities, 38*(1), 61-82.

Chang, C. C., & Lin, C. J. (2010). *LIBSVM — A library for support vector machines (Version 3.0)*. Retrieved May 18, 2011, from http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Churches, T., Christen, P., Lim, K., & Zhu, J. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, *2*(9). Retrieved May 18, 2011, from http://www.biomedcentral.com/1472-6947/2/9.

CiteSeer^X. (2011). *About CiteSeer^X*. Retrieved May 18, 2011, from http://citeseer.ist.psu.edu/about/site

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. *Proceedings of the AAAI 6th International Workshop on Information Integration*, 32-37.

Digital Author Identifier (DAI). (2010). *DAI-Standard wiki*. Retrieved May 18, 2011, from http://www.surffoundation.nl/wiki/display/standards/DAI

DiLauro, T., Choudhury, G. S., Patton, M., Warner, J. W., & Brown, E. W. (2001). Automated name authority control and enhanced searching in the levy collection. *D-Lib Magazine*, *7*(4). Retrieved May 18, 2011, from http://www.dlib.org/dlib/april01/dilauro/04dilauro.html

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *TKDE*, *19*(1), 1-16.

French, J. C., Powell, A., & Schulman, E. (2000). Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, *51*(8), 774-786.

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. *Proceedings of the 4th ACM/IEEE-CS Joint Conference,* 296-305. Retrieved May 18, 2011, from http://clgiles.ist.psu.edu/papers/JCDL-2004-author-disambiguation.pdf

Han, H., Giles, L., Zha, H., & Xu, W. (2005a). A hierarchical Naïve Bayes mixture model for name disambiguation in author citations. *Proceedings of the 2005 ACM symposium on Applied computing,* 1065-1069. Retrieved May 18, 2011, from http://clgiles.ist.psu.edu/papers/SAC-2005-Naïve-Bayes-Mixture.pdf

Han, H., Giles, L., & Zha, H., (2005b). Name disambiguation in author citations using a K-way spectral clustering method. *Proceedings of ACM/IEEE Joint Conference on Digital Libraries,* 334-343. Retrieved May 18, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.9354&rep=rep1&type=pdf

Hernandez, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*(1), 9-37.

Hill, S., & Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations*, *5*(2), 179-184.

Huang, J., Ertekin., S., & Giles, C. L. (2006). Efficient name disambiguation for large scale databases. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 536-544.

International Standard Name Identifier (ISNI). (2010). *International Standard Name Identifier Draft ISO 27729*. Retrieved May 18, 2011, from http://www.isni.org/

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, *14*(5-7), 491-498.

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.

Koppel, M., Argamon, S., & Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401-412.

Lu, C., Xu, Y., & Geva, S. (2008). Web-Based Query Translation for English-Chinese CLIR. *Computational Linguistics and Chinese Language Processing, 13*(1), 61-90.

Malin, B., Airoldi, E., & Carley, K. M. (2005). A network analysis model for disambiguation of names in lists. *Computational and Mathematical Organization Theory*, *11*(2), 119-139.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw Hill.

Online Computer Library Center (OCLC). (2010). *WorldCat Identity Service*. Retrieved from http://orlabs.oclc.org/identities

People Australia. (2010). *People Australia Overview*. Retrieved May 18, 2011, from http://www.nla.gov.au/initiatives/peopleaustralia/index.html

Pereira, D. A., Ribeiro-Neto, B. A., Ziviani, N., Laender, A. H. F., Gonçalves, M. A., & Ferreira, A. A. (2009). Using web information for author name disambiguation. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 49-58.

ProQuest. (2010). *Scholar Universe*. Retrieved May 18, 2011, from http://www.scholaruniverse.com

Research Name Resolver. (2010). *NII Research Name Resolver*. Retrieved May 18, 2011, from http://rns.nii.ac.jp/;jsessionid=372CE9C69AF0745A1597C34DD3ACC420

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1-43.

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.

Tan, Y. F., Kan, M. Y., & Lee, D. (2006). Search engine driven author disambiguation. In G. Marchionini, M. L. Nelson, & C. C. Marshall (Eds.), *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 314-315.

Thomson Reuters. (2010). *Distinct Author Identification System*. Retrieved May 18, 2011, from http://scientific.thomsonreuters.com/support/faq/wok3new/dais/

Thomson Routers. (2011). *Journal Citation Reports*. Retrieved May 18, 2011, from http://www.isiwebofknowledge.com/

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discover*y from Data, *3*(3), 1-29.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, *56*(2), 140-158.

Wiley-Blackwell. (2010). *Author Service*. Retrieved May 18, 2011, from http://authorservices.wiley.com/bauthor/

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. College, & P. S. Kott (ed.), *Business Survey Methods* (pp.355-384). New York: J. Wiley.

Yang, K. H., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2007). *Extracting citation relationships from web documents for author disambiguation*. Retrieved May 18, 2011, 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. *Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries,* 185-196. Retrieved May 18, 2011, from http://www.iis.sinica.edu.tw/papers/hoho/7642-F.pdf

Appendix

Performance of 14 author name datasets measured in accuracy (%).

A. Gupta (572 bibliographic records, 26 distinct authors)

K-means				Naïve Bayes				SVM			
C	12.7	**CT**	11.8	**C**	36.5	**CT**	35.8	**C**	75.4	**CT**	78.4
CY	18.7	**CTY**	18.3	**CY**	33.0	**CTY**	36.3	**CY**	76.5	**CTY**	78.3
CP	20.4	**CTP**	20.2	**CP**	36.6	**CTP**	36.2	**CP**	73.2	**CTP**	76.7
CYP	21.3	**CTYP**	21.1	**CYP**	37.0	**CTYP**	34.7	**CYP**	72.4	**CTYP**	77.4
T	11.8	**TJ**	23.7	**T**	35.2	**TJ**	38.6	**T**	67.6	**TJ**	71.2
TY	18.3	**TJY**	23.7	**TY**	33.6	**TJY**	37.1	**TY**	67.6	**TJY**	73.6
TP	20.2	**TJP**	20.2	**TP**	33.7	**TJP**	37.7	**TP**	65.5	**TJP**	72.9
TYP	21.1	**TJYP**	22.0	**TYP**	34.8	**TJYP**	37.6	**TYP**	66.6	**TJYP**	73.8
J	25.3	**CJ**	18.7	**J**	42.9	**CJ**	40.0	**J**	57.8	**CJ**	76.7
JY	22.9	**CJY**	20.8	**JY**	43.8	**CJY**	42.0	**JY**	61.3	**CJY**	78.1
JP	24.6	**CJP**	20.2	**JP**	41.7	**CJP**	41.1	**JP**	56.3	**CJP**	74.3
JYP	23.7	**CJYP**	22.2	**JYP**	44.1	**CJYP**	42.0	**JYP**	59.8	**CJYP**	77.3
CTJ	19.9	**CTJP**	20.2	**CTJ**	37.7	**CTJP**	38.2	**CTJ**	78.4	**CTJP**	78.0
CTJY	23.7	**CTJYP**	22.0	**CTJY**	38.8	**CTJYP**	38.0	**CTJY**	79.0	**CTJYP**	77.6

A. Kumar (238 bibliographic records, 14 distinct authors)

K-means				Naïve Bayes				SVM			
C	17.6	**CT**	17.6	**C**	41.9	**CT**	42.9	**C**	64.0	**CT**	71.4
CY	26.8	**CTY**	27.7	**CY**	44.3	**CTY**	42.0	**CY**	62.6	**CTY**	69.5
CP	32.3	**CTP**	31.0	**CP**	43.6	**CTP**	42.8	**CP**	66.1	**CTP**	69.4
CYP	24.3	**CTYP**	28.1	**CYP**	45.0	**CTYP**	45.4	**CYP**	64.2	**CTYP**	70.6
T	17.2	**TJ**	22.2	**T**	42.5	**TJ**	46.9	**T**	69.6	**TJ**	73.4
TY	27.7	**TJY**	27.7	**TY**	43.2	**TJY**	45.8	**TY**	69.2	**TJY**	76.6
TP	31.0	**TJP**	30.6	**TP**	44.1	**TJP**	46.0	**TP**	68.0	**TJP**	76.7
TYP	28.1	**TJYP**	28.5	**TYP**	45.0	**TJYP**	47.5	**TYP**	68.8	**TJYP**	76.1
J	26.4	**CJ**	28.1	**J**	51.0	**CJ**	48.4	**J**	70.4	**CJ**	77.8
JY	26.8	**CJY**	27.3	**JY**	52.4	**CJY**	48.3	**JY**	65.2	**CJY**	73.6
JP	31.5	**CJP**	31.0	**JP**	51.4	**CJP**	48.3	**JP**	64.6	**CJP**	74.8
JYP	28.9	**CJYP**	28.5	**JYP**	51.2	**CJYP**	46.9	**JYP**	64.6	**CJYP**	75.7
CTJ	20.5	**CTJP**	30.6	**CTJ**	45.3	**CTJP**	44.8	**CTJ**	76.5	**CTJP**	75.2
CTJY	27.7	**CTJYP**	28.5	**CTJY**	45.6	**CTJYP**	45.0	**CTJY**	76.0	**CTJYP**	76.6

C. Chen (679 bibliographic records, 61 distinct authors)

K-means				**Naïve Bayes**				**SVM**			
C	12.5	**CT**	10.8	**C**	17.4	**CT**	15.5	**C**	65.7	**CT**	60.1
CY	15.7	**CTY**	12.2	**CY**	17.6	**CTY**	14.9	**CY**	64.8	**CTY**	62.9
CP	17.2	**CTP**	12.0	**CP**	17.7	**CTP**	15.2	**CP**	62.8	**CTP**	62.1
CYP	14.5	**CTYP**	12.9	**CYP**	18.2	**CTYP**	14.8	**CYP**	60.9	**CTYP**	63.3
T	12.6	**TJ**	16.6	**T**	13.6	**TJ**	16.5	**T**	53.7	**TJ**	58.4
TY	12.0	**TJY**	15.7	**TY**	15.0	**TJY**	18.3	**TY**	51.6	**TJY**	60.0
TP	11.1	**TJP**	15.6	**TP**	14.0	**TJP**	17.5	**TP**	52.0	**TJP**	57.8
TYP	13.8	**TJYP**	14.4	**TYP**	16.1	**TJYP**	17.2	**TYP**	51.7	**TJYP**	58.9
J	23.7	**CJ**	17.5	**J**	23.5	**CJ**	22.6	**J**	43.7	**CJ**	66.7
JY	16.9	**CJY**	15.0	**JY**	26.3	**CJY**	23.9	**JY**	43.9	**CJY**	66.7
JP	19.7	**CJP**	15.1	**JP**	24.3	**CJP**	22.4	**JP**	41.5	**CJP**	65.3
JYP	17.0	**CJYP**	13.5	**JYP**	25.9	**CJYP**	23.4	**JYP**	43.9	**CJYP**	66.7
CTJ	15.1	**CTJP**	14.2	**CTJ**	16.3	**CTJP**	18.1	**CTJ**	64.6	**CTJP**	65.4
CTJY	15.1	**CTJYP**	15.3	**CTJY**	17.9	**CTJYP**	18.3	**CTJY**	65.5	**CTJYP**	64.2

D. Johnson (347 bibliographic records, 15 distinct authors)

K-means				Naïve Bayes				SVM			
C	31.7	**CT**	15.5	**C**	50.9	**CT**	50.9	**C**	73.9	**CT**	76.2
CY	32.2	**CTY**	31.7	**CY**	52.4	**CTY**	51.0	**CY**	76.9	**CTY**	77.3
CP	27.0	**CTP**	32.5	**CP**	51.2	**CTP**	51.0	**CP**	71.5	**CTP**	76.1
CYP	25.9	**CTYP**	26.5	**CYP**	51.6	**CTYP**	50.7	**CYP**	72.7	**CTYP**	78.5
T	15.5	**TJ**	29.9	**T**	51.2	**TJ**	51.3	**T**	70.7	**TJ**	75.4
TY	31.4	**TJY**	29.6	**TY**	49.8	**TJY**	50.7	**TY**	73.5	**TJY**	77.3
TP	32.5	**TJP**	32.2	**TP**	51.3	**TJP**	50.1	**TP**	72.6	**TJP**	75.8
TYP	29.1	**TJYP**	26.8	**TYP**	50.5	**TJYP**	51.3	**TYP**	74.4	**TJYP**	77.7
J	32.5	**CJ**	25.3	**J**	52.0	**CJ**	51.1	**J**	69.0	**CJ**	80.9
JY	34.8	**CJY**	30.8	**JY**	52.7	**CJY**	51.0	**JY**	67.9	**CJY**	79.5
JP	36.3	**CJP**	33.1	**JP**	52.3	**CJP**	49.8	**JP**	66.4	**CJP**	79.5
JYP	27.0	**CJYP**	26.5	**JYP**	54.6	**CJYP**	50.9	**JYP**	69.1	**CJYP**	79.7
CTJ	29.9	**CTJP**	32.8	**CTJ**	50.9	**CTJP**	50.7	**CTJ**	77.6	**CTJP**	78.7
CTJY	29.6	**CTJYP**	26.8	**CTJY**	50.4	**CTJYP**	49.8	**CTJY**	80.5	**CTJYP**	77.3

J. Lee (1270 bibliographic records, 99 distinct authors)

K-means				Naïve Bayes				SVM			
C	0.5	**CT**	0.2	**C**	12.5	**CT**	11.5	**C**	68.1	**CT**	70.4
CY	9.6	**CTY**	10.5	**CY**	11.9	**CTY**	9.3	**CY**	67.5	**CTY**	69.5
CP	11.6	**CTP**	11.2	**CP**	12.3	**CTP**	11.1	**CP**	64.6	**CTP**	69.3
CYP	11	**CTYP**	11.8	**CYP**	11.7	**CTYP**	11.4	**CYP**	63.7	**CTYP**	69.1
T	0.2	**TJ**	16.9	**T**	10.7	**TJ**	14.9	**T**	59	**TJ**	65.2
TY	9.7	**TJY**	15.9	**TY**	10.7	**TJY**	14.2	**TY**	60.5	**TJY**	65.1
TP	10.7	**TJP**	14	**TP**	11.5	**TJP**	12.5	**TP**	59.2	**TJP**	64.1
TYP	11.6	**TJYP**	11.4	**TYP**	10.8	**TJYP**	14.3	**TYP**	59.2	**TJYP**	63.5
J	18.3	**CJ**	16.8	**J**	18.6	**CJ**	16.1	**J**	47.6	**CJ**	69
JY	15.5	**CJY**	15.5	**JY**	18.7	**CJY**	16.8	**JY**	47.5	**CJY**	70.3
JP	16.4	**CJP**	12.9	**JP**	19.3	**CJP**	13	**JP**	46.3	**CJP**	69.7
JYP	13.3	**CJYP**	12.5	**JYP**	18.7	**CJYP**	16.3	**JYP**	45.8	**CJYP**	70
CTJ	16.2	**CTJP**	14.4	**CTJ**	13.6	**CTJP**	13.8	**CTJ**	73.2	**CTJP**	72.1
CTJY	14.8	**CTJYP**	12	**CTJY**	14.4	**CTJYP**	14.1	**CTJY**	72.5	**CTJYP**	72.2

J. Martin (103 bibliographic records, 15 distinct authors)

K-means				Naïve Bayes				SVM			
C	36.8	**CT**	21.3	**C**	15.9	**CT**	27.9	**C**	50.5	**CT**	47.4
CY	40.7	**CTY**	29.1	**CY**	28.3	**CTY**	32.6	**CY**	49.3	**CTY**	50.5
CP	36.8	**CTP**	23.3	**CP**	24.3	**CTP**	27.1	**CP**	43.0	**CTP**	48.0
CYP	32.0	**CTYP**	27.1	**CYP**	27.0	**CTYP**	21.8	**CYP**	45.2	**CTYP**	54.7
T	10.6	**TJ**	35.9	**T**	17.2	**TJ**	37.1	**T**	42.8	**TJ**	60.9
TY	26.2	**TJY**	30.9	**TY**	29.1	**TJY**	37.3	**TY**	49.0	**TJY**	62.7
TP	21.3	**TJP**	23.3	**TP**	22.9	**TJP**	36.3	**TP**	42.6	**TJP**	58.6
TYP	27.1	**TJYP**	32.0	**TYP**	22.2	**TJYP**	44.4	**TYP**	46.1	**TJYP**	66.1
J	44.6	**CJ**	36.8	**J**	47.0	**CJ**	40.5	**J**	56.3	**CJ**	62.3
JY	39.8	**CJY**	33.0	**JY**	45.3	**CJY**	40.4	**JY**	61.3	**CJY**	65.6
JP	33.9	**CJP**	30.0	**JP**	45.3	**CJP**	44.1	**JP**	50.7	**CJP**	61.7
JYP	37.8	**CJYP**	37.8	**JYP**	46.0	**CJYP**	41.6	**JYP**	54.9	**CJYP**	61.3
CTJ	36.8	**CTJP**	28.1	**CTJ**	38.8	**CTJP**	34.8	**CTJ**	60.1	**CTJP**	62.6
CTJY	31.0	**CTJYP**	34.9	**CTJY**	37.0	**CTJYP**	38.6	**CTJY**	62.8	**CTJYP**	68.3

J. Robinson (168 bibliographic records, 12 distinct authors)

K-means				Naïve Bayes				SVM			
C	41	**CT**	25	**C**	40.9	**CT**	34.4	**C**	69.0	**CT**	73.5
CY	33.3	**CTY**	26.7	**CY**	41.6	**CTY**	32.3	**CY**	65.8	**CTY**	75.3
CP	30.9	**CTP**	26.1	**CP**	40.2	**CTP**	32.4	**CP**	64.3	**CTP**	68.4
CYP	33.9	**CTYP**	30.3	**CYP**	43.9	**CTYP**	32.3	**CYP**	63.4	**CTYP**	75.2
T	14.2	**TJ**	24.4	**T**	33.0	**TJ**	37.7	**T**	55.5	**TJ**	68.5
TY	26.7	**TJY**	30.3	**TY**	33.9	**TJY**	38.7	**TY**	58.2	**TJY**	70.9
TP	24.4	**TJP**	29.1	**TP**	33.1	**TJP**	38.3	**TP**	58.1	**TJP**	68.7
TYP	30.3	**TJYP**	30.3	**TYP**	33.3	**TJYP**	42.2	**TYP**	60.3	**TJYP**	72.2
J	26.7	**CJ**	27.3	**J**	44.3	**CJ**	43.5	**J**	66.9	**CJ**	73.6
JY	30.9	**CJY**	29.1	**JY**	47.0	**CJY**	44.1	**JY**	62.5	**CJY**	72.0
JP	29.1	**CJP**	29.7	**JP**	47.2	**CJP**	45.0	**JP**	60.8	**CJP**	74.3
JYP	30.3	**CJYP**	35.1	**JYP**	47.3	**CJYP**	45.5	**JYP**	64.4	**CJYP**	72.0
CTJ	30.3	**CTJP**	30.3	**CTJ**	35.4	**CTJP**	37.6	**CTJ**	73.4	**CTJP**	76.3
CTJY	32.7	**CTJYP**	32.1	**CTJY**	37.6	**CTJYP**	40.7	**CTJY**	77.0	**CTJYP**	75.9

J. Smith (872 bibliographic records, 29 distinct authors)

K-means				Naïve Bayes				SVM			
C	15.3	**CT**	14.1	**C**	61.3	**CT**	54.3	**C**	80.2	**CT**	85.2
CY	31.9	**CTY**	25.1	**CY**	63.8	**CTY**	56.1	**CY**	77.3	**CTY**	84.8
CP	29.0	**CTP**	24.4	**CP**	61.9	**CTP**	55.9	**CP**	77.7	**CTP**	85.2
CYP	21.7	**CTYP**	20.1	**CYP**	64.7	**CTYP**	56.0	**CYP**	76.3	**CTYP**	85.7
T	14.1	**TJ**	17.6	**T**	42.2	**TJ**	61.3	**T**	74.4	**TJ**	83.2
TY	22.4	**TJY**	25.2	**TY**	45.4	**TJY**	62.5	**TY**	75.0	**TJY**	84.6
TP	24.4	**TJP**	23.6	**TP**	44.7	**TJP**	60.9	**TP**	72.4	**TJP**	83.0
TYP	19.6	**TJYP**	19.1	**TYP**	46.5	**TJYP**	61.5	**TYP**	74.4	**TJYP**	84.2
J	20.4	**CJ**	27.5	**J**	61.9	**CJ**	67.3	**J**	76.1	**CJ**	86.6
JY	21.5	**CJY**	24.3	**JY**	62.4	**CJY**	69.2	**JY**	76.4	**CJY**	85.8
JP	22.7	**CJP**	21.1	**JP**	63.0	**CJP**	67.5	**JP**	75.7	**CJP**	85.4
JYP	18.0	**CJYP**	19.6	**JYP**	62.5	**CJYP**	69.1	**JYP**	78.0	**CJYP**	85.6
CTJ	20.9	**CTJP**	23.3	**CTJ**	64.3	**CTJP**	64.2	**CTJ**	89.3	**CTJP**	88.4
CTJY	24.6	**CTJYP**	19.4	**CTJY**	63.7	**CTJYP**	65.7	**CTJY**	88.3	**CTJYP**	88.6

K. Tanaka (267 bibliographic records, 10 distinct authors)

K-means				Naïve Bayes				SVM			
C	18.1	**CT**	18.4	**C**	61.8	**CT**	60.0	**C**	83.4	**CT**	83.8
CY	34.7	**CTY**	35.8	**CY**	63.6	**CTY**	61.1	**CY**	82.4	**CTY**	86.4
CP	28.2	**CTP**	30.4	**CP**	60.9	**CTP**	59.7	**CP**	81.2	**CTP**	85.1
CYP	29.3	**CTYP**	23.5	**CYP**	63.5	**CTYP**	61.2	**CYP**	80.3	**CTYP**	84.8
T	18.4	**TJ**	21.3	**T**	54.8	**TJ**	62.5	**T**	78.5	**TJ**	84.6
TY	34.0	**TJY**	26.4	**TY**	58.6	**TJY**	65.0	**TY**	80.0	**TJY**	87.6
TP	30.4	**TJP**	30.4	**TP**	57.0	**TJP**	62.5	**TP**	77.7	**TJP**	84.4
TYP	29.3	**TJYP**	25.7	**TYP**	55.1	**TJYP**	63.4	**TYP**	80.8	**TJYP**	86.1
J	23.1	**CJ**	20.6	**J**	65.4	**CJ**	68.9	**J**	75.4	**CJ**	87.0
JY	28.9	**CJY**	28.6	**JY**	65.1	**CJY**	68.0	**JY**	74.4	**CJY**	89.5
JP	30.7	**CJP**	29.7	**JP**	65.2	**CJP**	69.3	**JP**	73.9	**CJP**	88.3
JYP	27.8	**CJYP**	25.3	**JYP**	66.3	**CJYP**	66.4	**JYP**	75.6	**CJYP**	86.5
CTJ	23.5	**CTJP**	31.1	**CTJ**	62.2	**CTJP**	65.8	**CTJ**	90.4	**CTJP**	87.1
CTJY	26.0	**CTJYP**	26.8	**CTJY**	64.1	**CTJYP**	63.6	**CTJY**	89.3	**CTJYP**	87.4

M. Brown (146 bibliographic records, 13 distinct authors)

K-means				Naïve Bayes				SVM			
C	30.1	**CT**	19.1	**C**	51.4	**CT**	38.3	**C**	72.5	**CT**	69.0
CY	37.6	**CTY**	36.9	**CY**	51.2	**CTY**	38.2	**CY**	71.7	**CTY**	72.3
CP	24.6	**CTP**	21.2	**CP**	45.9	**CTP**	38.0	**CP**	72.0	**CTP**	72.1
CYP	35.6	**CTYP**	39.7	**CYP**	48.3	**CTYP**	38.0	**CYP**	68.2	**CTYP**	72.6
T	15.0	**TJ**	23.2	**T**	30.8	**TJ**	36.0	**T**	66.0	**TJ**	67.8
TY	36.3	**TJY**	36.3	**TY**	34.0	**TJY**	40.2	**TY**	70.5	**TJY**	73.3
TP	21.2	**TJP**	25.3	**TP**	33.2	**TJP**	36.8	**TP**	66.8	**TJP**	70.6
TYP	39.7	**TJYP**	39.7	**TYP**	33.7	**TJYP**	40.8	**TYP**	63.8	**TJYP**	70.4
J	27.3	**CJ**	28.0	**J**	41.4	**CJ**	42.9	**J**	63.7	**CJ**	71.4
JY	36.9	**CJY**	36.3	**JY**	40.3	**CJY**	43.9	**JY**	60.6	**CJY**	71.7
JP	23.2	**CJP**	22.6	**JP**	42.8	**CJP**	49.0	**JP**	59.4	**CJP**	70.1
JYP	26.3	**CJYP**	39.7	**JYP**	48.1	**CJYP**	46.6	**JYP**	64.9	**CJYP**	76.2
CTJ	18.4	**CTJP**	24.6	**CTJ**	33.6	**CTJP**	46.5	**CTJ**	76.9	**CTJP**	76.2
CTJY	36.3	**CTJYP**	39.7	**CTJY**	45.3	**CTJYP**	43.1	**CTJY**	75.9	**CTJYP**	73.2

M. Jones (247 bibliographic records, 13 distinct authors)

K-means				Naïve Bayes				SVM			
C	38.0	**CT**	19.8	**C**	39.1	**CT**	44.6	**C**	60.1	**CT**	71.4
CY	37.6	**CTY**	26.3	**CY**	43.6	**CTY**	45.9	**CY**	60.7	**CTY**	69.5
CP	24.2	**CTP**	19.0	**CP**	46.7	**CTP**	48.3	**CP**	57.2	**CTP**	72.3
CYP	24.2	**CTYP**	21.4	**CYP**	46.1	**CTYP**	47.6	**CYP**	55.7	**CTYP**	71.6
T	15.7	**TJ**	22.6	**T**	45.1	**TJ**	54.2	**T**	65.0	**TJ**	79.8
TY	22.6	**TJY**	24.6	**TY**	47.6	**TJY**	51.1	**TY**	65.7	**TJY**	78.3
TP	19.4	**TJP**	21.0	**TP**	41.6	**TJP**	54.3	**TP**	65.3	**TJP**	79.3
TYP	23.4	**TJYP**	27.5	**TYP**	45.1	**TJYP**	53.9	**TYP**	66.2	**TJYP**	77.5
J	19.8	**CJ**	19.4	**J**	56.8	**CJ**	58.7	**J**	74.6	**CJ**	77.3
JY	26.3	**CJY**	25.1	**JY**	58.8	**CJY**	55.3	**JY**	74.3	**CJY**	77.9
JP	21.0	**CJP**	22.6	**JP**	58.8	**CJP**	54.8	**JP**	70.7	**CJP**	78.2
JYP	24.2	**CJYP**	24.2	**JYP**	57.1	**CJYP**	58.9	**JYP**	7.04	**CJYP**	78.8
CTJ	24.2	**CTJP**	21.4	**CTJ**	55.4	**CTJP**	55.6	**CTJ**	80.1	**CTJP**	81.5
CTJY	24.6	**CTJYP**	27.5	**CTJY**	55.5	**CTJYP**	54.6	**CTJY**	77.9	**CTJYP**	80.2

M. Miller (384 bibliographic records, 12 distinct authors)

K-means				Naïve Bayes				SVM			
C	18.4	**CT**	18.4	**C**	75.7	**CT**	66.7	**C**	84.4	**CT**	88.1
CY	43.4	**CTY**	42.9	**CY**	76.4	**CTY**	69.8	**CY**	85.8	**CTY**	86.6
CP	28.1	**CTP**	28.6	**CP**	75.8	**CTP**	68.3	**CP**	83.5	**CTP**	89.8
CYP	35.6	**CTYP**	35.6	**CYP**	77.5	**CTYP**	68.7	**CYP**	81.8	**CTYP**	88.7
T	18.4	**TJ**	18.4	**T**	58.8	**TJ**	61.4	**T**	84.9	**TJ**	85.8
TY	42.9	**TJY**	42.9	**TY**	58.0	**TJY**	60.7	**TY**	84.1	**TJY**	88.4
TP	28.6	**TJP**	25.7	**TP**	60.9	**TJP**	63.7	**TP**	85.0	**TJP**	87.8
TYP	35.6	**TJYP**	35.6	**TYP**	59.9	**TJYP**	62.1	**TYP**	84.6	**TJYP**	88.6
J	18.7	**CJ**	18.4	**J**	74.4	**CJ**	78.8	**J**	87.4	**CJ**	91.1
JY	44.7	**CJY**	42.9	**JY**	72.9	**CJY**	79.8	**JY**	87.0	**CJY**	90.7
JP	26.0	**CJP**	26.5	**JP**	74.6	**CJP**	79.2	**JP**	84.5	**CJP**	89.9
JYP	38.0	**CJYP**	35.6	**JYP**	74.3	**CJYP**	79.3	**JYP**	87.6	**CJYP**	90.2
CTJ	18.4	**CTJP**	25.7	**CTJ**	72.5	**CTJP**	67.5	**CTJ**	89.9	**CTJP**	91.1
CTJY	42.9	**CTJYP**	35.6	**CTJY**	67.0	**CTJYP**	69.0	**CTJY**	88.6	**CTJYP**	89.9

S. Lee (1260 bibliographic records, 84 distinct authors)

K-means				Naïve Bayes				SVM			
C	4.7	**CT**	1.4	**C**	15.2	**CT**	14.9	**C**	69.5	**CT**	67.8
CY	8.2	**CTY**	13.6	**CY**	15.6	**CTY**	15.0	**CY**	68.6	**CTY**	66.6
CP	14.1	**CTP**	12.7	**CP**	15.2	**CTP**	14.9	**CP**	66.9	**CTP**	64.9
CYP	15.3	**CTYP**	14.5	**CYP**	15.5	**CTYP**	15.1	**CYP**	66.3	**CTYP**	67.1
T	1.4	**TJ**	17.6	**T**	14.7	**TJ**	17.0	**T**	58.9	**TJ**	67.2
TY	12.9	**TJY**	16.7	**TY**	14.8	**TJY**	17.1	**TY**	59.2	**TJY**	66.5
TP	11.5	**TJP**	15.7	**TP**	14.9	**TJP**	17.4	**TP**	58.5	**TJP**	67.0
TYP	14.6	**TJYP**	15.6	**TYP**	14.8	**TJYP**	17.0	**TYP**	58.9	**TJYP**	66.8
J	26.5	**CJ**	18.6	**J**	26.1	**CJ**	18.7	**J**	53.3	**CJ**	74.0
JY	19.7	**CJY**	15.5	**JY**	26.8	**CJY**	19.0	**JY**	55.1	**CJY**	72.4
JP	18.4	**CJP**	16.5	**JP**	26.5	**CJP**	18.8	**JP**	53.3	**CJP**	72.7
JYP	18.9	**CJYP**	16.5	**JYP**	27.2	**CJYP**	18.6	**JYP**	55.7	**CJYP**	73.2
CTJ	17.1	**CTJP**	15.0	**CTJ**	15.9	**CTJP**	16.2	**CTJ**	71.5	**CTJP**	72.0
CTJY	16.3	**CTJYP**	14.2	**CTJY**	15.8	**CTJYP**	16.0	**CTJY**	70.6	**CTJYP**	72.4

Y. Chen (1168 bibliographic records, 71 distinct authors)

K-means				Naïve Bayes				SVM			
C	0.7	**CT**	0.5	**C**	23.2	**CT**	22.2	**C**	70.8	**CT**	68.6
CY	19.9	**CTY**	16.1	**CY**	23.9	**CTY**	22.6	**CY**	69.3	**CTY**	70.4
CP	17.2	**CTP**	15.7	**CP**	23.8	**CTP**	22.2	**CP**	65.4	**CTP**	70.2
CYP	18.1	**CTYP**	16.5	**CYP**	24.8	**CTYP**	22.3	**CYP**	67.3	**CTYP**	72.4
T	0.5	**TJ**	12.5	**T**	21.8	**TJ**	26.6	**T**	62.6	**TJ**	68.0
TY	16.8	**TJY**	17.8	**TY**	22.1	**TJY**	27.0	**TY**	64.8	**TJY**	70.4
TP	15.0	**TJP**	12.1	**TP**	22.6	**TJP**	27.1	**TP**	63.6	**TJP**	67.8
TYP	16.0	**TJYP**	14.5	**TYP**	22.9	**TJYP**	27.2	**TYP**	64.0	**TJYP**	68.4
J	16.4	**CJ**	14.8	**J**	30.9	**CJ**	27.7	**J**	53.0	**CJ**	72.7
JY	18.6	**CJY**	17.2	**JY**	31.1	**CJY**	28.0	**JY**	55.4	**CJY**	74.6
JP	15.1	**CJP**	12.0	**JP**	31.5	**CJP**	28.3	**JP**	52.1	**CJP**	72.8
JYP	15.7	**CJYP**	14.1	**JYP**	31.8	**CJYP**	29.0	**JYP**	54.0	**CJYP**	74.0
CTJ	15.6	**CTJP**	13.8	**CTJ**	25.9	**CTJP**	26.3	**CTJ**	71.8	**CTJP**	72.6
CTJY	18.7	**CTJYP**	14.5	**CTJY**	26.2	**CTJYP**	25.9	**CTJY**	73.9	**CTJYP**	73.4

教育資料與圖書館學　49 : 2 (Winter 2011) : 215-240

研究論文

書目資料中著者姓名歧義性之解析ψ

陳光華*

教授
臺灣大學圖書資訊學系
E-mail: khchen@ntu.edu.tw

謝其男

研究生
臺灣大學圖書資訊學系
E-mail: r97126004@ntu.edu.tw

摘要

目前網際網路已經快速地累積大量的學術資訊，使用者經常會面臨到著者歧異性的問題，使得對同名著者群的解析成為一項重要的研究課題。相較於前人研究，本研究充分應用文獻書目資料僅有的資訊，而不使用書目資訊以外的資訊。本研究探討「共同著者姓名(C)」、「文獻題名(T)」、「期刊題名(J)」、「出版年(Y)」、「頁數(P)」等五項特徵資訊，其中「出版年」與「頁數」從未有其他研究使用過。本研究使用監督式學習方法(Naïve Bayes與SVM)與非監督式分類方法(K-means)，探討28項不同的特徵資訊組合。研究發現「期刊題名(J)」與「共同作者(C)」是特別有效的特徵資訊；J在三種方法皆有很好的表現，C則是在SVM方法有很好的效用。「出版年(Y)」與「頁數(P)」在與其他特徵資訊的組合明顯地提升歧義性解析的正確率，兩者以「出版年(Y)」的輔助效果較為突出(平均提升2.5%)。在前人研究中經常被使用的特徵資訊組合「CTJ」並不一定能取得最佳的正確率，而JYP、JY、CJ等特徵組合亦能達到最佳的正確率。最後比較資料集的規模與複雜度的實驗結果發現，規模較大複雜度較高的資料集的準確率低了10%，顯示當測試的資料集日益龐雜時，完全倚靠書目資料難以提供令人滿意的辨識效果。顯現在未來研究中，若要有效地解決人名歧異性之問題，除了充分使用書目資料的各項特徵，仍須使用適當的外部資訊。

關鍵詞：著者歧義性，書目資料，機器學習

ψ 本文部分內容曾發表於《教育資料與圖書館學》40週年國際學術研討會，2011年3月7-8日。

* 本文主要作者兼通訊作者。