國立臺灣大學

圖書資訊學研究所

碩士學位論文

書目資料中著者姓名歧義性之解析

Ambiguity Resolution of Author Names for Bibliographic Data

研 究 生：謝其男

指導教授：陳光華博士

中華民國一〇〇年七月

Ambiguity Resolution of Author Names for Bibliographic Data

Author： Chi-Nan Hsieh

Advisor： Kuang-hua Chen, Ph. D.

A Master Thesis Submitted to

Department of Library and Information Science

National Taiwan University

July 2011

謝辭

呼～終於到了能夠撰寫謝辭的時候了…

在這本又輕又薄的論文裡頭其實集結了許多師長前輩、親朋好友的貢獻。首先，當然要最感謝這三年以來恩師陳光華教授的指導與照顧，我自從一年級上過老師的課後就決定將來一定要跟著老師來進行研究，從一開始對自然語言處理完全沒有概念，到最後可以去自行開發與執行，都要多虧恩師這些年來數不盡的耐心教導與指點。接著，也十分感謝在論文計畫書口試與學位口試時所擔任委員的兩位老師：唐牧群教授與黃乾綱教授，提供了許多具有建設性建議與觀念的傳授，讓我的研究得以順利進行與完成。

除了三位口試委員外，我也很感激系上的每位老師，在修課時提供了許多之後在自己論文進行時的隱性大補丸。慕萱老師對學術水準的期許與嚴謹、明德老師對資訊科技與資訊組織的宏觀思維、寶媛老師對資訊呈現上的使用者導向、珊如老師在研究方法上的完整詮釋、書梅老師對於各種書目療法的推廣、奇秀老師的批判思考與社會科學應有的細膩觀察力、雪華老師對未來圖書館走向的積極態度、文欽老師對圖書史學的重視等重要觀念，都給予我許多學術研究上重要的刺激與啟蒙。

在我碩士生活的日子中，學校與系所的助教朋友們也都提供了很多求學上的支援。非常感謝宜玲、盈達、逸晴、佩民以及喻淳等每一位助教所給予我的無微不至的照顧，系辦、系圖與系資的大家都是我求學生涯中不可或缺的避風港。另外也要感謝第七研究室的宜芳學姐、雅蓁學姐與家豪，總是都不留餘力地給我許多論文與生活上的建議與協助。此外也非常感謝校史館人文中心的宗銘學長、以及總是在系館裡忙上忙下、親切和藹的夏小姐給我論文進行上的幫忙與鼓勵。

由於在我論文實驗進行時缺乏跑資料的電腦，這時候才發現自己身邊的朋友一個個都願意義氣相挺，給我有第三年畢業的機會，感謝光華老師、盈達助教、世娟學姐、家豪、瑋妮、富任、瑞庭、佩瑜、佩瑜的哥哥、老爸跟老姐等人的熱情贊助。碩士生活的同學與朋友們也是我論文進行時的最佳防空洞，大家總是一團和樂的互相激勵、聊天與充電，超級感謝彥翔、建豪、瑋妮、家

虹、佳馨、馥蓉、思岑、立芳、瑋麟、欣怡、瑋安、凱傑、亞真、郁文、恬安、彥如等所有系上的同學們。在文獻蒐集時，大學部的林禹伸等學弟們也對我的論文貢獻良多，實至感謝。在論文的程式開發時，感謝大學同學佳伶提供我撰寫 Python 語言的重要書籍。在一、二年級外宿永和的日子裡，許多大學同學也經常來關心或聚餐，感謝秉修、竣榮、有崇、柏鈞、楚鈞、培軒、意晴、貝珊、琇婷、依紋、心儀等輔大的好同學們。還有畢業的驚奇四超人的拼股、秋刀魚與婷婷公主，感謝大家時時刻刻的互相鼓勵與出遊聚會。

還有我親愛的玩偶團們，總是睜大眼睛地陪伴在我身邊，感謝形影不離的團長猴子瑞可、最淑女的瑞可可、嫁去小魚家的熱可可、背包上的大雄、來自福岡的福可、大力士浩呆、傲嬌的漢娜等等，都是我每天好心情的來源呢！

最後，這份論文要獻給我最親的家人，爸爸、媽媽以及姊姊，你們都是我最重要的支柱，家裡頭的溫情永遠都是最好我的堡壘，提供我持續完成學業的動力。還有這一年來女友佩瑜的相伴與打氣，讓我能克服瓶頸、一步一步地將論文給結束，也才有了同時畢業的機會。

我想，在短短的文字裡是絕對繁不及述所有給我幫助過的人。因此，我決定引用陳之藩先生的經典名句：「要感謝的人太多了，那就謝天吧！」感謝老天爺讓我遇到你們大家，使我在學術的道路上並不孤獨，也才有現在眼前這本論文的誕生。

摘要

在檢索大量的學術資訊時，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的資訊進行辨識工作，且不使用書目資訊以外的資訊。因此，我們使用「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。本研究分別使用監督式學習方法與非監督式分類方法，探討總共 28 項不同的特徵資訊組合，分別對著者姓名歧義性解析的正確率。

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，其中「期刊題名（J）」無論在各種方法中都展現重要性，而「共同作者（C）」則主要在使用支持向量機（Support Vector Machine，SVM）方法時十分出色。另外，「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（約平均提升 2.5%），此外出版年與頁數對歧異性解析的影響效果在使用 K-means 分群方法時的特別明顯（約 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，透過不同分類方法發現其他特徵組合亦能達到最佳的正確率，如 JYP、JY、CJ 等特徵組合。最後根據資料集的規模與複雜度進行辨識結果的比較中發現，當測試的資料集日益龐雜時，僅倚靠引用文獻的書目資料則難以提供充足的辨識效果。顯現在未來研究中，若要有效地解決人名歧異性之問題，必須從書目資料的資訊向外與其他資訊進行連結與對應，以獲取更明確的作者特徵。

關鍵詞：著者歧義性、 書目資料、機器學習

Abstract

In order to solve name ambiguity when retrieving academic information, researches on author identification are indispensable. With comparison to previous works, this study attempts to address this 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 extracted from bibliographic data and will be used to disambiguate author names in this work. 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 approaches, and feature C is mainly outstanding in SVM. In addition, feature year (Y) and feature number of pages (P) obviously enhance accuracy rate while they accompanied with various feature combination(s), and the average improvement rate of inclusion with feature Y is more significant than feature P. However, it is significant that 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 (+0.15% in average).

It is also shown that the performance of feature combination CTJ used traditionally is not superior to JYP, and the performance of feature combinations CJY, JY and J are also very effective in three methods. Finally, it is found that the accuracy of disambiguation on larger datasets is 10% inferior to the smaller ones, which indicated the limitation and deficiency of the performance achieved by bibliographic data in this “numerous and jumbled” real world. Consequently, it is a promising trend in the future to build an intellectual mechanism to map other information onto bibliographic information accurately in order to get sufficient information for author disambiguation.

Keywords: Author Disambiguation, Bibliographic Data, Machine Learning

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

In general, names seem helpful in identifying a person with great ease. However, with widespread use of digital information in Internet era, name ambiguity problems have commonly occurred. The name 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 way towards the problem. International Standard Organization (ISO, 2010) has established International Standard Name Identifier (ISNI) 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 provided to handle this problem. The first approach is building supplementary identification functions 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 the ambiguous name 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 service 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. Nevertheless, 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 issues mentioned above show 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.

1.2 Objectives of Research

Two objectives of this study are: 1) To explore how the performance can be achieved by using complete bibliographic data only, which is composed of authors, article titles, journal titles, publication date, and number of pages and 2) To investigate how the performance can be influenced with consideration of publication date and number of pages, which have never been discussed before.

1.3 Restriction of Research

In order to compare our results to previous works, the datasets of this study are followed by Han et al. (2005). Therefore, the coverage of data collection in our experiments is only "bibliographic data" instead of considering outside resources, such as web information (Yang et al., 2007, 2008).

1.4 Definition of Terms

1.4.1 Bibliographic data

Bibliographic data can provide reference information to readers. In general, bibliographic data contain: author(s), title, edition, publisher, publication place, publication date, number of pages, etc. According to our datasets which mainly composed by journal or conference paper, bibliographic data in this article include: author(s), title (or article title), journal title, publication date, and number of pages.

1.4.2 Ambiguity Resolution

Ambiguity resolution is the mathematical process/algorithm for determining ambiguities. Having a determined initial integer ambiguity value for each satellite, the integrated carrier phase measurement can be used as a precise distance measurement between the receiver and satellites (Navman Glossary, 2011). In this study, the targets of ambiguity resolution are authors sharing the same name, and the disambiguation work is used for measurement between these authors.

Chapter 2

Literature Review

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 research issue. Finally, machine learning approaches are described, and the methods in our experiment also introduced.

2.1 Name Disambiguation

The problem of name ambiguity originates in a much broader issue: identity uncertainty and the study of pioneers in the area called "record linkage" by Fellegi and Sunter (1969). They developed a statistical model to process multiple records in one or more 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 and 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 and 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 machine 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., 2005).

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 emphasized the distance between strings (Torvik et al., 2005), and others focused on 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.

There are several types of name disambiguation studies below, and show the current status of this issue.

a) 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).

b) 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), and so on.

c) 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, 2005; 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).

d) The application on the records in multimedia database, 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).

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

2.2 Ambiguity Resolution for Author

As mentioned above, several research task forces devoted themselves to author name disambiguation for different purposes. “CiteSeer” is a famous digital library service developed by Steve Lawrence, Lee Giles and Kurt Bollacker (CiteSeer, n.d.). 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. Thus, the identification of name or author identification is a key work. 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, 2005; Huang, Ertekin, & Giles, 2006). Later studies managed to apply additional features of data, such as the first page of the paper. In addition, many different unsupervised learning models were used, e.g., probabilistic latent semantic analysis and latent Dirichlet allocation (LDA) (Song et al., 2007).

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 characteristic of their works includes 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.

"Author-ity" is an author name disambiguation system for MEDLINE using the features of co-authors, journal titles, article titles, subject headings, language, affiliations and author name. That is to say, some features not available in bibliographic data were used in this system. Probabilistic model is used for implementation of this system and the performance is claimed achieving the recall of 98.8% (Torvik, 2009).

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 outside resources are taken into account for delivering satisfactory performance as well, such as full-text articles and information from web pages. Nevertheless, 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.

In Han's studies (Han et al, 2004, 2005), 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. (2005) 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	Best 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) 2) 73.3% (Bayes)
Han et al. (2005)	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. (2005)	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)	91.3% (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)	92.5% (25% better than Han's K-way)

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.

2.3 Machine Learning

Like the approaches for Data mining and text mining, machine learning are used in our disambiguation experiments. In general, machine learning methods include two types: Supervised learning methods and unsupervised learning methods. The types and introductions of both machine learning methods are described in this section.

2.3.1 Supervised Learning Methods

Supervised learning methods include two-steps: training and classification. In the former step, a model would be built by training data set composed of samples which is selected from total population randomly, and class labels are pre-assigned to each training data of the learning process. Then, in the second step, the model is used for classification. The predictive accuracy of the model is estimated by using test set (also randomly selected). The accuracy is considered as the percentage of test samples correctly classified. If the accuracy is acceptable, the model will apply to classify unknown data to their appropriate classes. Otherwise, the model needs modification until it meets an acceptable level of classification accuracy. Major techniques of supervised learning methods involve:

- Bayesian Classification: Bayesian classifiers are statistical classifiers based on Bayes theorem in probability theory. Bayes theorem is defined as:

  $$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

  Let X be a data sample whose class label is unknown. Let H be some hypothesis such that sample X belongs to class C. The probability that H holds on data sample X is the posterior probability defined as $P(H \mid X)$. In contrast, $P(H)$ is the prior probability of H , which is independent of X . Similarly, $P(X \mid H)$ is the posterior probability of X conditioned on H. $P(X)$ is the prior probability of X. In additional, *Naive Bayes classifier* is an instance of a particular kind of Bayes classifier (Gale et al., 1992), and it assume class conditional independence. In the other words, a feature value for a given class is independent of the values of the other feature. Mitchell (1997) also pointed out that Naïve Bayes is widely used in machine learning duo to its efficiency and its ability to combine evidence from a large number of features. Therefore, Naïve Bayes classifier is used in our disambiguation work for authors.

- Decision Trees: Decision Tree Classifiers (DTC's for short) are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, and etc. (Safavian & Landgrebe, 1991). A decision tree is constructed from a training set, which consists of objects. Each object is completely described by a set of attributes and a class label. Attributes can have ordered (e.g., real) or unordered (e.g., Boolean) values. A decision tree contains zero or more *internal* nodes and one or more *leaf* nodes. All internal nodes have two or more *child* nodes. All *internal* nodes contain *splits*, which test the value of an expression of the attributes. *Arcs* from an internal node *t* to its children are labeled with distinct outcomes of the test at *t*. Each leaf node has a class label associated with it (Murthy, 1998).

- K-Nearest Neighbor: The k-nearest-neighbor classifier (KNNC for short) is one of the most basic classifiers for pattern recognition or data classification. The principle of this method is based on the intuitive concept that data points of the same class should be closer in the feature space. As a result, for a given data point x of unknown class, we can simply compute the distance between x and all the data points in the training data, and assign the class determined by the K nearest points of x. Due to the simplicity of KNNC, it is often used as a baseline method in comparison with other sophisticated approaches in pattern recognition (Jang, 2011).

- Support Vector Machine: The support vector machine (SVM for short) is a new machine technique used for classifier. SVM is introduced by Vapnik (1995) in his work on structure risk minimization, and it attempts to construct a hyperplane partitioning two sets of observations, where each observation is an element of a low-dimensional space. An interesting characteristic of these models is the volume of data, which can lead to quadratic programs with between 10 and 100 million variables and, if written explicitly, a dense Q matrix (Ferris & Munson, 2002). In this study, we also conduct SVM in disambiguation work by LibSVM tool (Chang & Lin, 2010).

2.3.2 Unsupervised Learning Methods

In contrast to supervised learning, the object class labels are not pre-given in unsupervised learning methods. Clustering (or clustering analysis), one common form of unsupervised learning, is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering analysis has a wide range of applications, including information retrieval, image processing, business transaction analysis, and pattern recognition. Two major types of clustering analysis are introduced as follows.

- Hierarchical clustering: Hierarchical methods construct a hierarchical decomposition of the given set of data objects using either an agglomerative (also called “bottoms-up”) or a divisive (also called “top-down”) approach.

Agglomerative strategies start at the bottom and at each level recursively merge a selected pair of clusters into a single cluster. This produces a grouping at the next higher level with one less cluster. The pair chosen for merging consists of the two groups with the smallest intergroup dissimilarity. Divisive methods start at the top and at each level recursively split one of the existing clusters at that level into two new clusters. The split is chosen to produce two new groups with the largest between-group dissimilarity (Hastie, 2011).

- Partitional clustering: Partitioning methods typically create an initial partition, which is then refined using iterative relocation techniques to improve the partitioning. Iterative relocation technique improves the partitioning by moving objects from one group to another. K-means clustering is one of most common partitional clustering methods, and aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (Yang et al, 1999). Thus, K-means clustering method is also employed in our experiment as unsupervised learning approach for author disambiguation work.

After the overview of the machine learning approaches above, different characteristics of supervised and unsupervised methods are found. And, the previous studies in Table 1 show that two types of machine learning were all employed. Therefore, both of supervised and unsupervised approaches are conducted in our experiment. The detail of methods we used is described in next chapter.

Chapter 3

Research Design

In order to investigate different factors, e.g., feature combinations, learning methods, and scalability of datasets, many resources are used and arranged in this study. The research framework is shown in Figure 1. The procedure consists of data collection, data processing, model learning, and performance evaluation. The following subsections explain these stages. In addition, feature encoding, feature combinations, and feature weightings are discussed in detail.

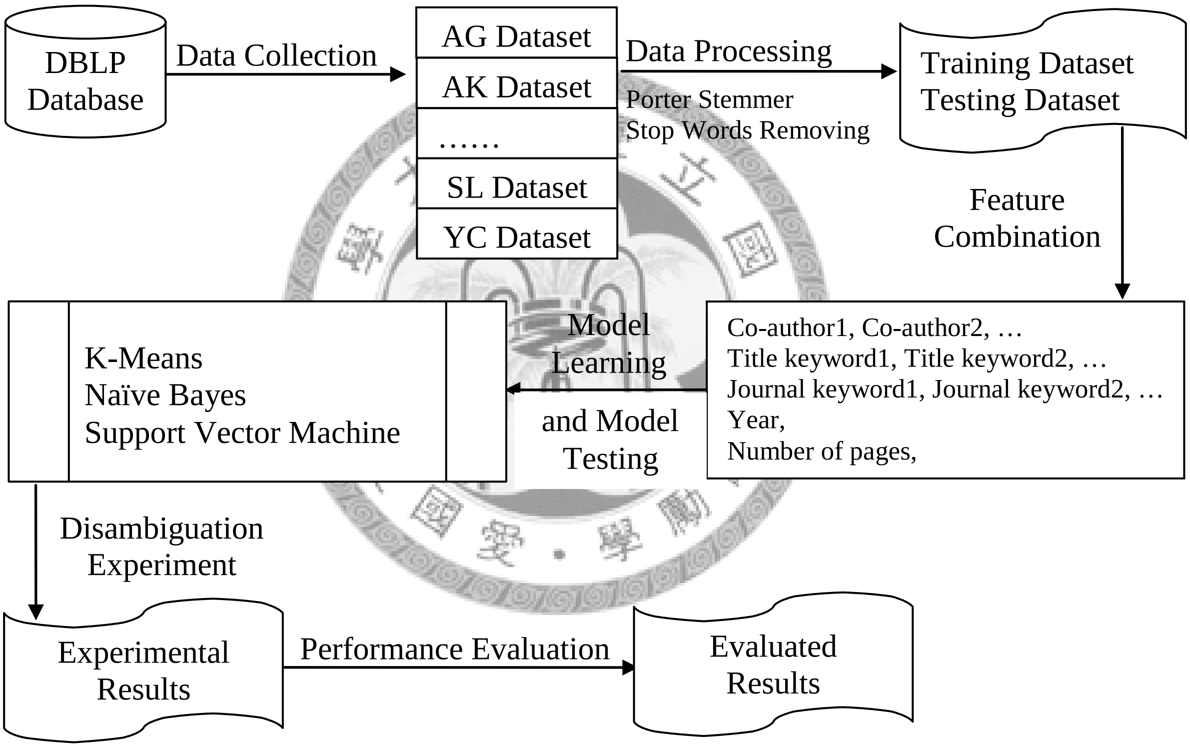


Figure 1: Research Procedure

3.1 Data Collection

The datasets employed in this study was the same DBLP datasets constructed by Han et al. (2005), 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 Han et al. (2005) 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 14 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**

3.2 Feature Combinations

The purpose of this study focuses on performance of complete combinations of various features (i.e. authors, article titles, journal titles, date, and number of pages) in bibliographic 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 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 previously "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

3.3 Data Processing

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 weightings, 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.

3.4 Machine Learning

After data processing, each bibliographic record is transferred into each 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 with cluster module using 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. The cluster of the maximum is first regarded as one class, and the second cluster is regarded as the other class and so on.

3.5 Performance Evaluation

Like Han et al. (2005) 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 then calculated 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.

3.6 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.

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 (More than 5,000 words)	**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 (More than 4,000 words)	**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	

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). 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.

Chapter 4

Experimental Results

In this study, 14 author names of DBLP datasets are examined (see Table 2 above). Each feature combination is investigated, and the effects of features Y and P are discussed. 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.

4.1 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 comparisons of one-feature (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 satisfactory. In two-feature (CT, TJ and CJ) experiment, feature CJ scored 78.5% of the lead in the comparisons of two-feature (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).

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, but it is not the case in Support Vector Machine (SVM for short). And, the performance of feature C is generally more desired than feature J in SVM. 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. Nevertheless, feature CTJ leads other feature combinations only 7 times in the 42 times of comparisons of the best accuracy, and the 6 times among the 7 times which feature CTJ obtained the lead were conducted by SVM. As a result, when features C, T, and J are used for disambiguation at the same time, the combination cannot necessarily ensure the best performance.

As above, 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). For example, feature C outperformed feature J or T, and it is believed “Hybrid scheme” (feature CTJ called in Han’s paper) was outstanding. However, the methods they conducted were only supervised, and the datasets they used were not the same as the experiment used in the study (see Table 1).

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

Support Vector Machine (SVM) - Continuing										
Rank of Single-Feature				Rank of Two-Feature				Best Accuracy	
	C	T	J		CT	TJ	CJ		CTJ
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									

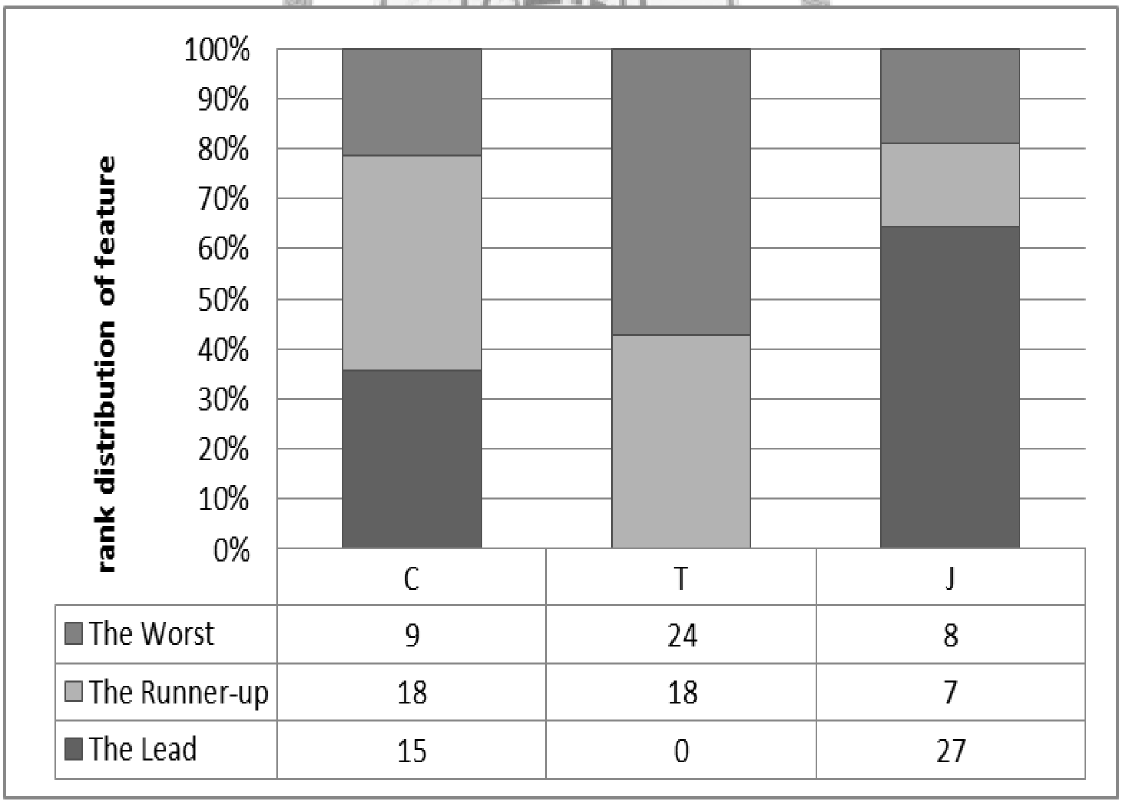


Figure 2: Rank Comparisons of Single Feature

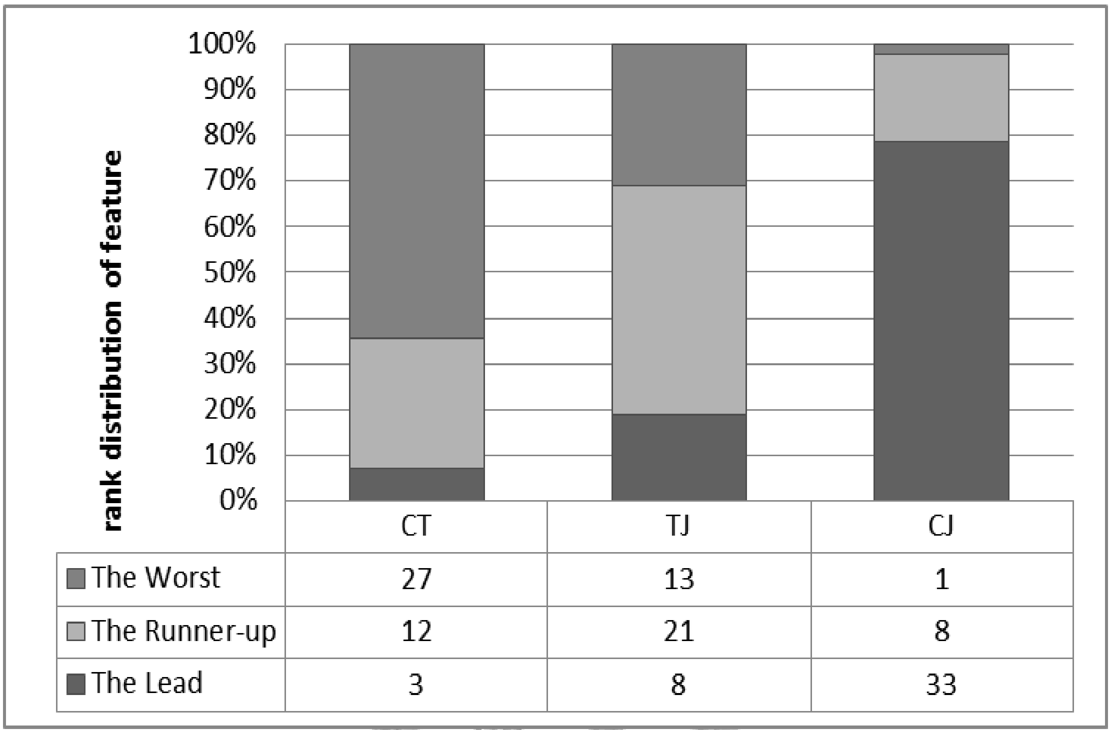


Figure 3: Rank Comparisons of Two Features

4.2 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. 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 using features Y and P is better than the previous one in general.

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 accuracy 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 accuracy 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 accuracy 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 accuracy 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).

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

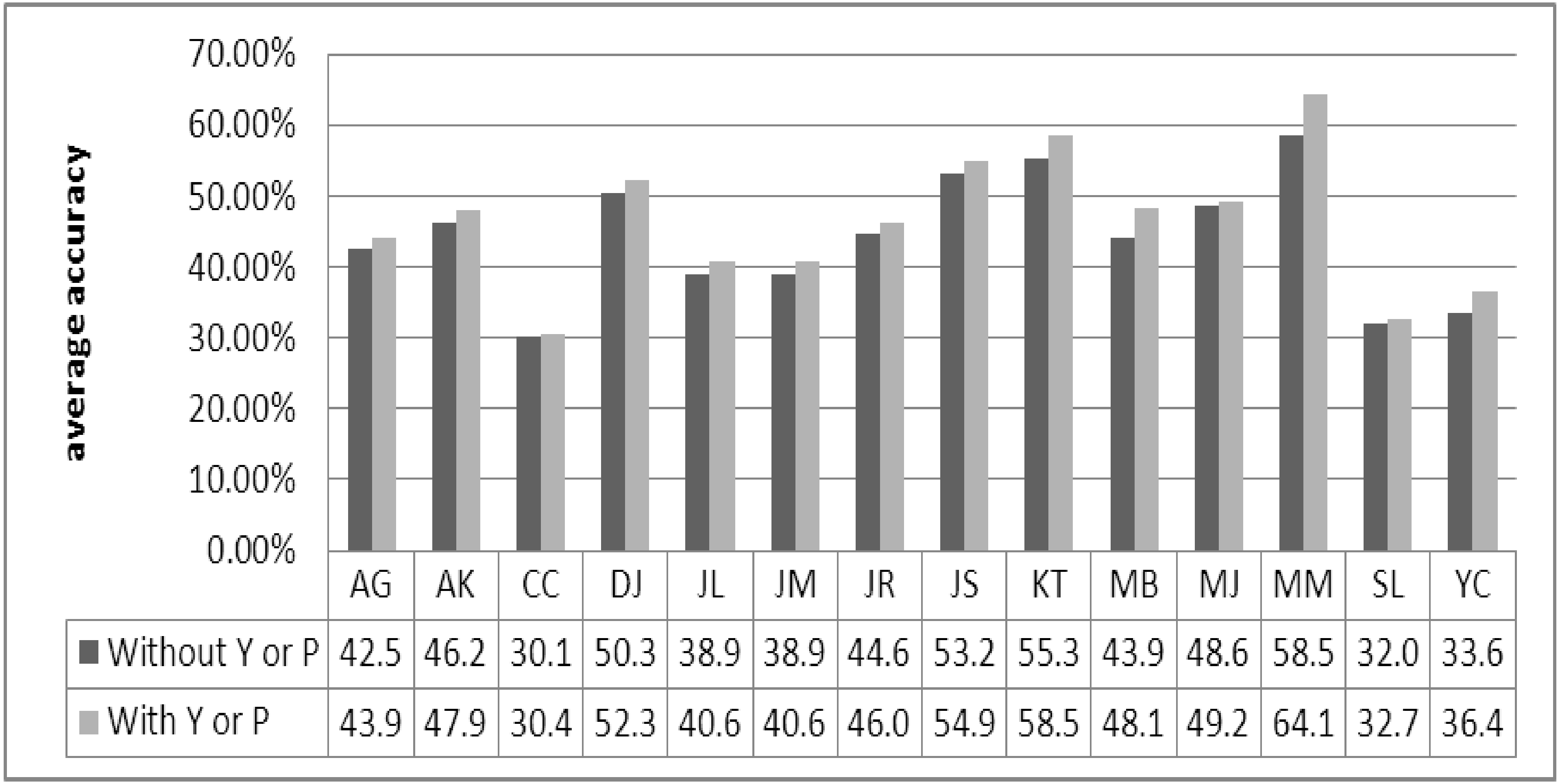
	KM			**NB**			**SVM**		
	Y	P	YP	Y	P	YP	Y	P	YP
AG	2.89	3.16	4.99	0.47	0.63	0.60	0.97	-1.43	0.30
AK	-1.24	9.53	8.81	0.07	-0.13	0.17	-1.57	-0.77	0.69
CC	0.43	0.41	0.13	0.10	-0.11	1.19	0.89	0.17	0.24
DJ	5.69	5.69	1.19	0.11	0.01	0.41	1.21	0.59	2.27
JL	3.20	3.16	2.07	-0.27	-0.63	-0.09	0.06	-1.03	-1.29
JM	0.86	-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**

From the findings shown above, it is found that feature Y and feature YP delivered positive performance in our datasets. In addition, the inclusion of feature P also produced positive effects, but the influence is not obvious. However, it is significant that 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 (+0.15% in average). Please refer to Figure 6. It is shown that feature Y and feature P could enhance significant performance in K-means clustering, but not obviously in Naïve Bayes and SVM. In the experiment by 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 by 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 rate in various unsupervised approaches in future studies.

4.3 Complexity of Datasets

According to the scale of datasets, the datasets are divided into two groups: Group A and Group B. Group A contains the complicated dataset (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 dataset (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.

As shown in Figure 4, 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 more negative 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 indeed. In other words, it is easier to increase ambiguity in larger datasets like the complexity in the real world.



	AG	AK	CC	DJ	JL	JM	JR	JS	KT	MB	MJ	MM	SL	YC
Without Y or P	42.5	46.2	30.1	50.3	38.9	38.9	44.6	53.2	55.3	43.9	48.6	58.5	32.0	33.6
With Y or P	43.9	47.9	30.4	52.3	40.6	40.6	46.0	54.9	58.5	48.1	49.2	64.1	32.7	36.4

Figure 4: The Comparison using with/out Features Y and P (Average in Three Methods)

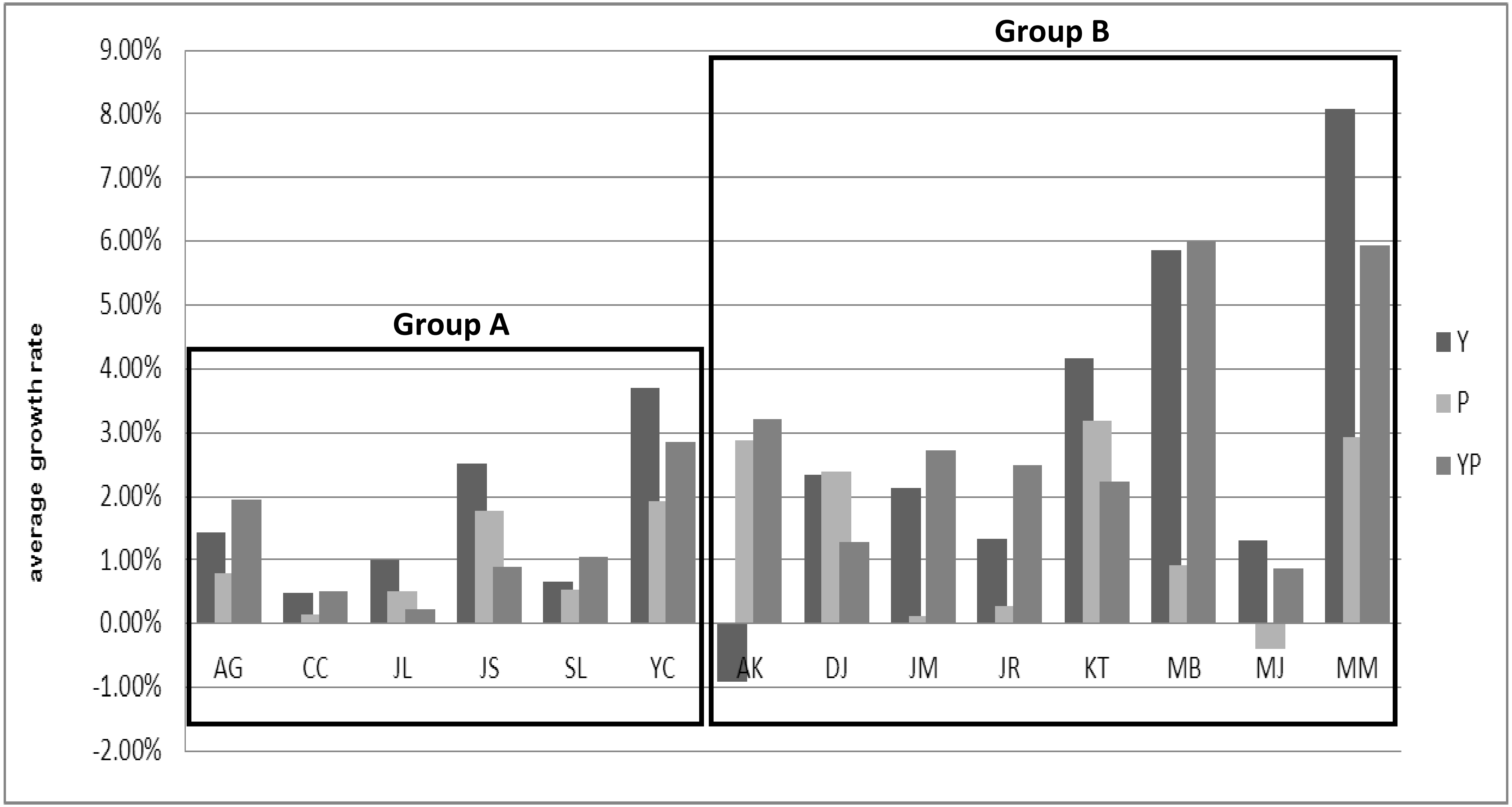


Figure 5: Average Improvement Rate using Features Y and P (Average in Three Methods)

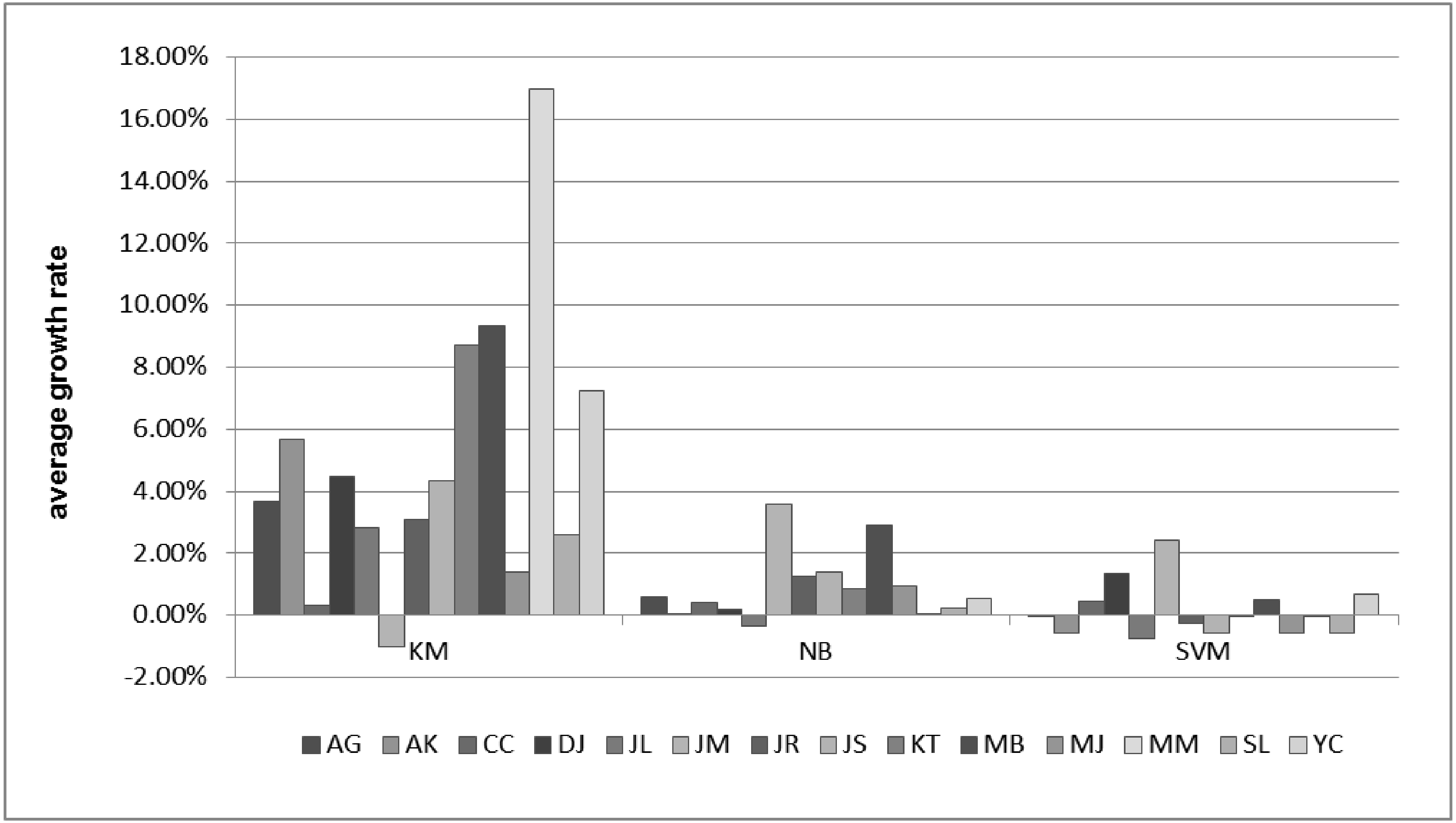


Figure 6: Improvement Accuracy Rate using Features Y and P in Different Methods (Average of Y, P and YP)

4.4 Top One Feature Combinations

Feature combinations achieving the best accuracy are explored in this part. 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.

There are 14 feature combinations in 18 top 1 feature combinations in Table 7 with inclusion of feature Y or feature P. That means features Y and P have their roles in author name disambiguation even though they were not ever considered before. In addition, feature J accounted for 77.7% of top 1 feature combinations, and feature C for 64.4% subsequently. Please refer to Figure 8. As Section 4.4 mentioned, it is found that when feature C and feature combination CTJ achieved outperformance is employed by SVM method.

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

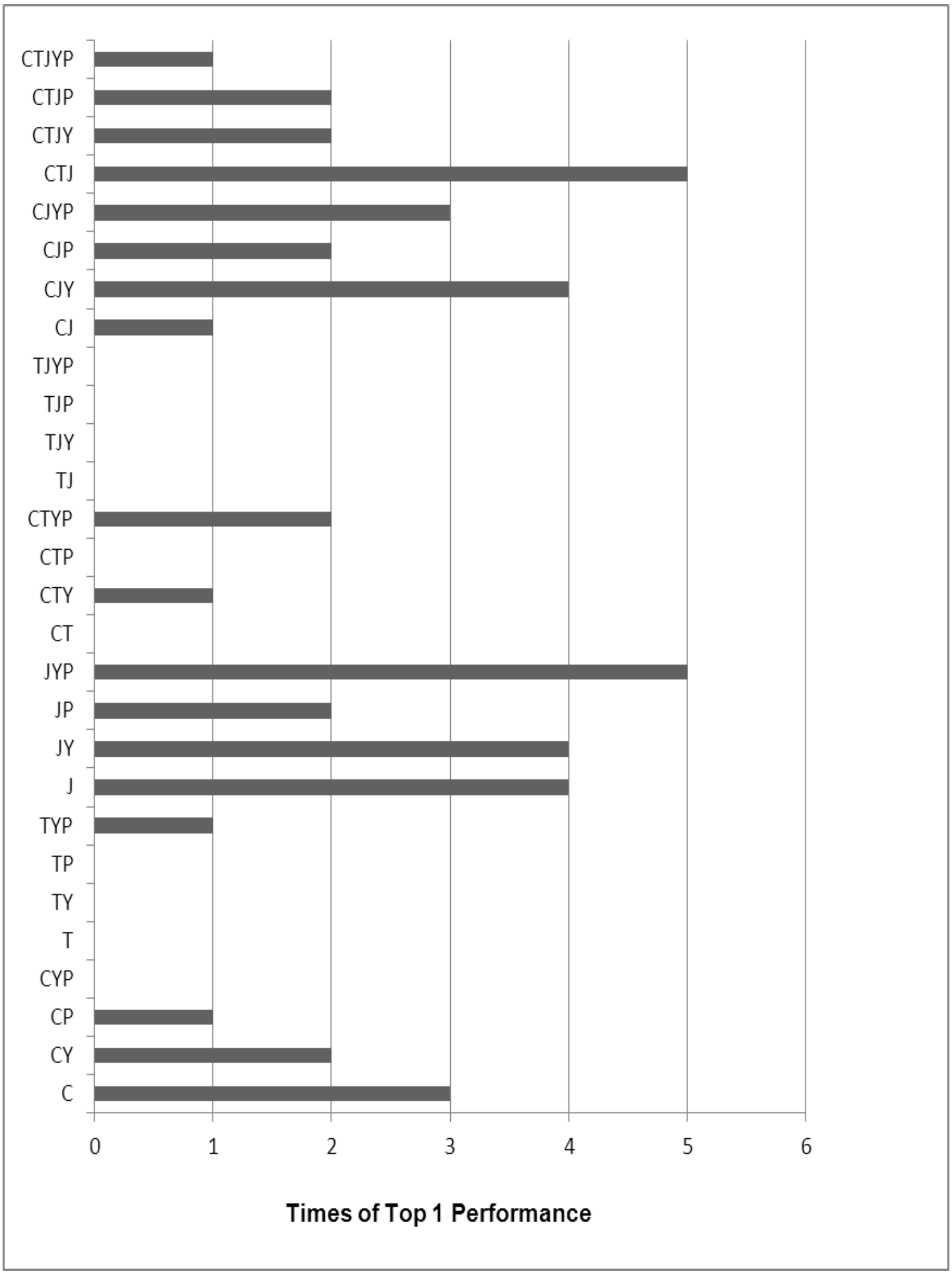


Figure 7: Top 1 Distribution of Feature Combinations

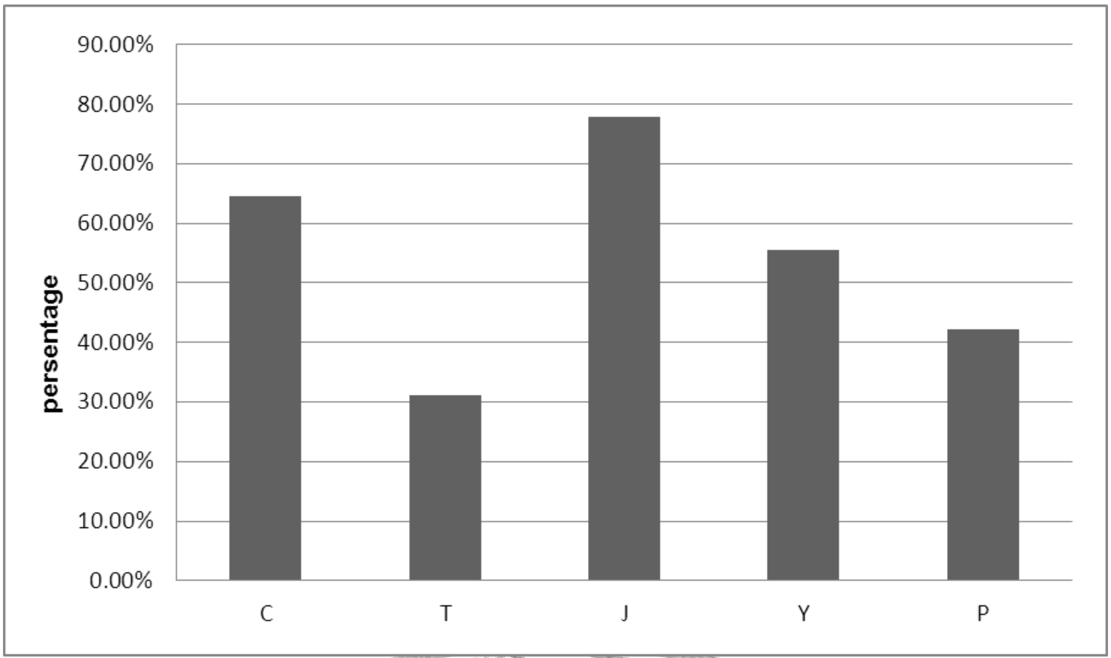


Figure 8: Percentage of Features in Top 1 Feature Combinations

Chapter 5

Conclusions and Suggestions

Finally, research conclusions are organized from the findings of the thesis in this section, and some research prospects are suggested for future studies.

5.1 Conclusions

According to the experimental results, some conclusions are taking shape and described as follows:

- Feature combination CTJ cannot necessarily ensure the best performance: In previous works, this common feature combination was usually regarded as a normal scheme, and the focus of studies often contributed to the designs of algorithm or the impacts of new resource. It is few to pay much attention to conduct a serial of different feature combinations repeatedly on author disambiguation. In this thesis, it is shown that the performance of feature combination JYP is not inferior to CTJ, and the performance of feature combinations CJY, JY and J are also outstanding in general. Therefore, it is known that the best feature combination on author disambiguation is mainly contributed by the combinations of features C and J. Additionally, the inclusion of features Y and P can substantially enhance the performance as well
- The inclusion of features Y and number of pages P exhibits positive influence on disambiguation: The average improvement rates of the inclusion of features Y are 2.44%, 1.29% in feature P, and 2.30% in YP. As Section 4.2 mentioned, the impacts of inclusions by features Y and P are significant in K-means clustering (about 5% accuracy of improvement). However, the influence of them is not obvious in Naïve Bayes and Support Vector Machine. It seems feasible to explore whether the feature Y and P could efficiently enhance accuracy rate in various “unsupervised” approaches in future studies. In addition, the setting for year and number of pages ought to depend on the character of datasets in order to respond to different datasets. For example, the setting for number of pages of journals in the datasets which consists of the citation records in humanity or social science should be more than 17 (used in our experiment).

- Various feature combinations have different effects on author name disambiguation while using different clustering or learning methods: It is found that the performance of feature combination J and JYP in K-means clustering and Naïve Bayes Model is as excellent as that of feature combination C and CTJ in SVM. Moreover, as the previous findings suggested, average improvement rate of using features Y and P in K-means (4.98%) is markedly better than Naïve Bayes (0.90%), but the growth rate in SVM is not effective at all (0.15%). In other words, it is shown that the selection of bibliographic feature information for author disambiguation work in the future could be applied according to the approaches of classification or clustering.
- The scale of datasets probably takes effects on the disambiguation work owing to the different complexity of datasets: The accuracy of disambiguation on larger datasets usually is lower than that of the smaller ones, and the effectiveness is not obvious while adding features Y and P. Although this causality is inferable, it clearly pointed out the limitation of the performance achieved by bibliographic data only. As a consequence, it can be expected that how to effectually recommend outer resource (ex: web information) is a critical issue in the future studies of name or author disambiguation in order to supplement additional accuracy rates from feature information.

5.2 Suggestions for Future Studies

The objectives of this study are to investigate effects of complete combinations of features contained in bibliographic data without resort to outside information. The current conclusion casts light on the usage of publication date and number of pages. There are some suggestions for further studies in author disambiguation, even though several feature combinations and different tools for classification or clustering had been implemented in this study.

- Exploration of performance of feature combinations from different dataset (rather than DBLP datasets only): 14 datasets in this study were composed of DBLP database by Han (2005). However, subject area of citations in DBLP database is only “Computer Science”. Therefore, it is worthy to explore whether the performance of feature will be influenced by authors/people from different disciplines.

- More complicated approaches to classification or clustering: Three existing tools (ex: K-means clustering model by Python, Naive Bayes by NLTK, SVM by LibSVM) were used in this study, but they are not very “tailor-made” in disambiguation work when comparing with Latent Dirichlet allocation (LDA) by Song et al. (2007) or 3-way and high-order simultaneous comparisons by McCallum et al. (2007). So, more sophisticated algorithms can be implemented in future studies.
- Enhancement of performance by various outside resources: It is challenging to completely solve author ambiguity by bibliographic information “only”, because bibliographic information in disambiguation work still generates a certain degree of “noise”. In this way, the performance cannot achieve acceptable standard (more than 90%) in general. Thus, it is a promising trend in the future to build an intellectual mechanism to map outside information onto bibliographic information accurately in order to get sufficient information for disambiguation.

References

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

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

Chang, C. C. & 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/

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.

CiteSeer (n.d.). About CiteSeer^{X}. Retrieved Jan. 31, 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. In: *Proceedings of the AAAI 6 th International Workshop on Information Integration* on the Web, 32-37.

Digital Author Identifier (DAI). (2009). *DAI-Standard wiki*. Retrieved Oct. 4, 2010, 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).

Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007). Duplicate record detection: A survey. *TKDE*, 19(1), p1–16.

Ferris, M. & Munson, T. (2002). Interior-point methods for massive support vector machines. *SIAM Journal on Optimization 13 (3)*: 783–804.

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

Gale, W. A., Church, K. W. & Yarowsky, W. (1992). A method for disambiguation word senses in a large corpus. *Computers and the Humanities 26*: 415-439.

Han, H., Giles, L., Zha, H., (2005a). 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. (2005b). 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

Hastie, T., Tibshirani, R., Friedman, J. (2011). Hierarchical clustering. *The Elements of Statistical Learning (2nd ed.). New York: Springer*, 520–528.

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), p9–37.

Hill, S., & Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. *ACM SIGKDD Explorations*, 5, 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). (2009). *ISNI Draft ISO 27729*. Retrieved Oct. 4, 2010, Retrieved Nov. 27, 2009, from http://www.isni.org/

Jang, J. S. (2011). Data Clustering and Pattern Recognition. Retrieved Jan. 4, 2011, Retrieved Dec. 25, 2010, from http://mirlab.org/jang

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14, 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, 401-412.

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

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

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*.

Naveman. (2011). Naveman Glossary. Retrieved Jan. 4, 2011, Retrieved Dec. 25, 2010, from http://www.navmanmarine.net/

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

People Australia. (2010). *People Australia Overview*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, 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. In *Proc. of JCDL*, pp 49–58.

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

Research Name Resolver. (2010). *NII Research Name Resolver*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://rns.nii.ac.jp/;jsessionid=372CE9C69AF0745A1597C34DD3ACC420

Safavian, S. R., Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Systems Man Cybernet. 21*, 660-674.

Smalheiser, N. R., Torvik, V. I. (2009). Author Name Disambiguation. Chapter in *Annual Review of Information Science and Technology*, v.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 Reuter. (2009). *Distinct Author Identification System*. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://scientific.thomsonreuters.com/support/faq/wok3new/dais/

Thomson Routers. (2011). Journal Citation Reports. Retrieved Oct. 4, 2010, Retrieved Jan. 3, 2011, from http://www.isiwebofknowledge.com/

Torvik V. I, Smalheiser N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery* from Data, 3(3):11.

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, 140-158.

Wiley-Blackwell. (2010). Author Service. Retrieved Oct. 4, 2010, Retrieved Dec. 25, 2009, from http://authorservices.wiley.com/bauthor/

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox et al. (Eds.), *Business Survey Methods*, New York: J. Wiley, 355-384.

Yang, D. L., Chang, J. H., Huang, M. C. & Liu, J. S. (1999). An efficient K-means-based clustering algorithm. *In Proceedings of the 1st Asia-Pacific Conference on Intelligent Agent Technology*, 269-273.

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

Appendix

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

A. Gupta (572 bibliographic records, 26 different 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			CTJ	37.7			CTJ	78.4		
CTJY	23.7			CTJY	38.8			CTJY	79.0		
CTJP	20.2			CTJP	38.2			CTJP	78.0		
CTJYP	22.0			CTJYP	38.0			CTJYP	77.6		

A. Kumar (238 bibliographic records, 14 different 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			CTJ	45.3			CTJ	76.5		
CTJY	27.7			CTJY	45.6			CTJY	76.0		
CTJP	30.6			CTJP	44.8			CTJP	75.2		
CTJYP	28.5			CTJYP	45.0			CTJYP	76.6		

C. Chen (679 bibliographic records, 61 different 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			CTJ	16.3			CTJ	64.6		
CTJY	15.1			CTJY	17.9			CTJY	65.5		
CTJP	14.2			CTJP	18.1			CTJP	65.4		
CTJYP	15.3			CTJYP	18.3			CTJYP	64.2		

D. Johnson (347 bibliographic records, 15 different 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			CTJ	50.9			CTJ	77.6		
CTJY	29.6			CTJY	50.4			CTJY	80.5		
CTJP	32.8			CTJP	50.7			CTJP	78.7		
CTJYP	26.8			CTJYP	49.8			CTJYP	77.3		

J. Lee (1270 bibliographic records, 99 different 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			CTJ	13.6			CTJ	73.2		
CTJY	14.8			CTJY	14.4			CTJY	72.5		
CTJP	14.4			CTJP	13.8			CTJP	72.1		
CTJYP	12			CTJYP	14.1			CTJYP	72.2		

J. Martin (103 bibliographic records, 15 different 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			CTJ	38.8			CTJ	60.1		
CTJY	31.0			CTJY	37.0			CTJY	62.8		
CTJP	28.1			CTJP	34.8			CTJP	62.6		
CTJYP	34.9			CTJYP	38.6			CTJYP	68.3		

J. Robinson (168 bibliographic records, 12 different authors)											
K-means				Naïve Bayes				SVM			
C	41	CT	25	C	40.9	CT	34.4	C	69	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	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	CJY	44.1	JY	62.5	CJY	72
JP	29.1	CJP	29.7	JP	47.2	CJP	45	JP	60.8	CJP	74.3
JYP	30.3	CJYP	35.1	JYP	47.3	CJYP	45.5	JYP	64.4	CJYP	72
CTJ	30.3			CTJ	35.4			CTJ	73.4		
CTJY	32.7			CTJY	37.6			CTJY	77		
CTJP	30.3			CTJP	37.6			CTJP	76.3		
CTJYP	32.1			CTJYP	40.7			CTJYP	75.9		

J. Smith (872 bibliographic records, 29 different 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	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	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	TJY	84.6
TP	24.4	TJP	23.6	TP	44.7	TJP	60.9	TP	72.4	TJP	83
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	CJP	67.5	JP	75.7	CJP	85.4
JYP	18	CJYP	19.6	JYP	62.5	CJYP	69.1	JYP	78	CJYP	85.6
CTJ	20.9			CTJ	64.3			CTJ	89.3		
CTJY	24.6			CTJY	63.7			CTJY	88.3		
CTJP	23.3			CTJP	64.2			CTJP	88.4		
CTJYP	19.4			CTJYP	65.7			CTJYP	88.6		

K. Tanaka (267 bibliographic records, 10 different authors)											
K-means				Naïve Bayes				SVM			
C	18.1	CT	18.4	C	61.8	CT	60	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	TJY	26.4	TY	58.6	TJY	65	TY	80	TJY	87.6
TP	30.4	TJP	30.4	TP	57	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
JY	28.9	CJY	28.6	JY	65.1	CJY	68	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			CTJ	62.2			CTJ	90.4		
CTJY	26			CTJY	64.1			CTJY	89.3		
CTJP	31.1			CTJP	65.8			CTJP	87.1		
CTJYP	26.8			CTJYP	63.6			CTJYP	87.4		

M. Brown (146 bibliographic records, 13 different authors)											
K-means				Naïve Bayes				SVM			
C	30.1	CT	19.1	C	51.4	CT	38.3	C	72.5	CT	69
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	CP	72	CTP	72.1
CYP	35.6	CTYP	39.7	CYP	48.3	CTYP	38	CYP	68.2	CTYP	72.6
T	15	TJ	23.2	T	30.8	TJ	36	T	66	TJ	67.8
TY	36.3	TJY	36.3	TY	34	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	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	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			CTJ	33.6			CTJ	76.9		
CTJY	36.3			CTJY	45.3			CTJY	75.9		
CTJP	24.6			CTJP	46.5			CTJP	76.2		
CTJYP	39.7			CTJYP	43.1			CTJYP	73.2		

M. Jones (247 bibliographic records, 13 different authors)											
K-means				Naïve Bayes				SVM			
C	38	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	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	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	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	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	74	CJYP	78.8
CTJ	24.2			CTJ	55.4			CTJ	80.1		
CTJY	24.6			CTJY	55.5			CTJY	77.9		
CTJP	21.4			CTJP	55.6			CTJP	81.5		
CTJYP	27.5			CTJYP	54.6			CTJYP	80.2		

M. Miller (384 bibliographic records, 12 different 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	TJY	60.7	TY	84.1	TJY	88.4
TP	28.6	TJP	25.7	TP	60.9	TJP	63.7	TP	85	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	CJY	90.7
JP	26	CJP	26.5	JP	74.6	CJP	79.2	JP	84.5	CJP	89.9
JYP	38	CJYP	35.6	JYP	74.3	CJYP	79.3	JYP	87.6	CJYP	90.2
CTJ	18.4			CTJ	72.5			CTJ	89.9		
CTJY	42.9			CTJY	67			CTJY	88.6		
CTJP	25.7			CTJP	67.5			CTJP	91.1		
CTJYP	35.6			CTJYP	69			CTJYP	89.9		

S. Lee (1260 bibliographic records, 84 different 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	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	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
TYP	14.6	TJYP	15.6	TYP	14.8	TJYP	17	TYP	58.9	TJYP	66.8
J	26.5	CJ	18.6	J	26.1	CJ	18.7	J	53.3	CJ	74
JY	19.7	CJY	15.5	JY	26.8	CJY	19	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			CTJ	15.9			CTJ	71.5		
CTJY	16.3			CTJY	15.8			CTJY	70.6		
CTJP	15			CTJP	16.2			CTJP	72		
CTJYP	14.2			CTJYP	16			CTJYP	72.4		

Y. Chen (1168 bibliographic records, 71 different 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
TY	16.8	TJY	17.8	TY	22.1	TJY	27	TY	64.8	TJY	70.4
TP	15	TJP	12.1	TP	22.6	TJP	27.1	TP	63.6	TJP	67.8
TYP	16	TJYP	14.5	TYP	22.9	TJYP	27.2	TYP	64	TJYP	68.4
J	16.4	CJ	14.8	J	30.9	CJ	27.7	J	53	CJ	72.7
JY	18.6	CJY	17.2	JY	31.1	CJY	28	JY	55.4	CJY	74.6
JP	15.1	CJP	12	JP	31.5	CJP	28.3	JP	52.1	CJP	72.8
JYP	15.7	CJYP	14.1	JYP	31.8	CJYP	29	JYP	54	CJYP	74
CTJ	15.6			CTJ	25.9			CTJ	71.8		
CTJY	18.7			CTJY	26.2			CTJY	73.9		
CTJP	13.8			CTJP	26.3			CTJP	72.6		
CTJYP	14.5			CTJYP	25.9			CTJYP	73.4		