

Towards Using Fewer Features for Text Classification

Yuan Yuan
Brisoft Information Technology Inc.
Beijing, China
yyyuan@gmail.com

Tianyang Gu
Brisoft Information Technology Inc.
Beijing, China
gty@brisoft.com.cn

Abstract—Text classification or categorization is a conventional classification problem applied to the text domain. In the cases when statistical classification methods are used, an important research issue is the selection of features from the training texts, each of which is hence treated as a feature vector. In this paper, we propose an approach for feature selection in text classification tasks, based on the exploit of external information that summarizes the text to be classified. In particular, we study the use of their *citation contexts* in the categorization of academic publications using the Naive Bayesian method. A series of experiments have been performed on a corpus of publications in Computer Science, based on which we observe that publication citation contexts can serve as a liable and effective source of feature selection. We also derive some useful hints on the reduction of feature number with a negligible affects on the accuracies.¹

I. INTRODUCTION

A. Background

Text classification (or categorization) is a conventional classification problem applied to the text domain. The general goal of a text classification task is to classify text content to one or more predefined categories, thus providing a way to organizing the text content. As the volume of text is continuously growing both on the Web and in corporate domains, text classification becomes critical not only from an academic point of view but also for industrial applications.

A number of statistical classification methods have been applied to text classification, such as Naive Bayesian, Bayesian Network, Support Vector Machines, Neural Network, and Decision Tree. A comparative evaluation of such text classification methods has been reported on various datasets like the Reuters corpus. As borrowed from the field of machine learning, most of these statistical classifiers usually treat a text document as a bag-of-words, thus represented by a feature vector, where each feature is one of the tokens appearing in the text.

One problem when using a statistical classifier to classify text documents is how to decide and extract the features from the text contents in a corpus. A number of methods have been proposed for this purpose. The simplest form of features are single stemmed or non-stemmed tokens, when the text uses a bag-of-words representation. To make use of the token

dependency and position information, it is suggested to use *phrasal features* consisting of more than one token [14], [3]. More sophisticated feature selection strategies utilize the Information Gain measures [10], [20].

B. Motivation and Overview

The extensive research on bibliographic citations has presented the usefulness of citation information to the tasks such as text retrieval, text summarization, and text classification [15], [13], [18], [12], [4]. It has been shown that the results also hold in the Web environment [1], [8], [19].

The application of bibliographic citations for text content analysis include two aspects: the citation structure and the citation content. In general, citation structures (or Web link structures) include information such as co-citation and bibliographic coupling, whereas citation contents refer to the *reference areas* in the citing document, where the observing document is cited [13]. In both bibliographic analysis and Web content analysis, citation contents can be further divided into different types according to their roles in the entire text [5], [12].

In this paper, we propose an approach for feature selection, which exploits some external information that summarizes the text to be classified. In particular, considering the many good properties presented by publication citations, we propose to use the *citation contexts* as a source of feature selection for the sake of text categorization by using the Naive Bayesian method. Our approach is inspired by the work of Glover et al., which proposes to use a “virtual document” of a Web page (consisting of the text of the citation to this page in different citing documents) for text classification [6]. The results of their experiments on Web pages (extracted from Yahoo! Categories) with less than 20 in-linked pages (obtained by Google) demonstrate very high accuracy of the proposed method.

In our approach, we see a text as a bag of words, and represent the text in a binary *feature vector*. The value of an item in the vector is 1 if the feature appears in the observing document, and 0 otherwise. Given the three different sources: text contents, citation contexts, and both, we use and compare three feature selection methods, namely *local-threshold* based method, *global-threshold* based method, and *most-frequent-term* based method, which take into account the *term frequency* in different ways.

¹The work presented in this paper was mostly done when the first author was working in Brisoft Inc. Yuan Yuan has now moved to: 1434 West Flournoy Street, IR, Chicago, IL 60607, USA (phone: 312-799-1854).

We perform a series of experiments on the effects of different factors (i.e., sources and methods) of the feature selection to the classification performances. Our experiments are carried out over the Cora dataset, which consists of a collection of academic publications in Computer Science. The Naive Bayesian method is chosen as our testing classifier.

We make the following two contributions in this paper:

- First, we justify the result of previous works that citation contexts are indeed a discriminative and descriptive representation of a text, thus providing a liable and effective source for feature selection.
- Second, we find out several unsophisticated ways to reduce the feature number while keeping a high level of classification accuracies. They include: using a low local threshold in the method of feature selection from both text and citations, using a high local threshold in the content-based feature selection, and using few most frequent terms extracted from the text content.

The rest of this paper is organized as follows. We describe the relevant work in Section II. Our main idea is presented in Section III. In Section IV, we describe the settings for our experiments and the results. Finally, we conclude in Section V.

II. RELATED WORK

The research on bibliographic citations has been long standing in the area of Bibliometrics [15], [18], [12], [4]. It has been testified that bibliographic citations are helpful to the content analysis for the sake of *associative document retrieval* [15]. More specifically, it is reported that the set of words extracted from a document, when supplemented by new words obtained from the bibliographic information, can provide a more accurate representation of document content and hence a more effective retrieval mechanism.

Web citations (or links), as a variant of bibliographic citations, have also been extensively studied on how they can help improve the performances of information retrieval, although Web citation analysis is not yet a replacement for the study of bibliographic citations [19]. Typical examples of applications of Web citations include: the PageRank algorithm that uses *link structure* and *anchor text* to improve the quality of results as well as the order in which the results are returned [1], the HITS algorithm for discovery of “authoritative” information sources on broad search topics [8], and some page ranking algorithms extended based on both algorithms [7], [17].

Another general application area of bibliographic citations is text summarization. As an example, Nanba and Okumura have proposed a method to construct a survey of an academic domain by analyzing the *reference areas* of multiple referring papers, in which references are categorized into different *reference types* (or citation types) based on lexical clues [13].

There already exists some work on using citation information for text classification in different ways. Continuing with their previous work, Nanba et al. discussed the classification of academic papers based on *topic similarity* that is calculated using either a text-based approach or a citation-based

approach [12]. The text-based approach utilizes different representative text such as title, abstract, full text, and two other special kinds of text (i.e., PURPOSES and METHODS of the paper). The citation-based approach uses citation information by two means: bibliographic coupling (i.e., two papers cite a common set of papers) without citation types and that with citation types. The latter means, in which a coupling paper will not be counted if their citation types do not coincide, has been proved most effective.

In the Web environment, Web link structures such as citation and bibliographic coupling have also been exploited in traditional content-based classification methods to improve the classification of Web collections [2]. Experiments on a Web directory show that the best result is achieved when the links from pages outside the directory are considered.

In addition to the use of link structures, Glover et al. has proposed to utilize the text near the citation in citing documents [6]. Results show that the text in citing documents, when available, often has greater discriminative and descriptive power than the text in the target document itself. In particular, the proposed approach works as follows: 1) extracting important features and training a full-text SVM classifier of web pages, 2) creating “virtual documents” from the anchor-text and inbound extended anchor-text, which is then used as a replacement for the full-text used by the original classifier, 3) combining the results to improve accuracy, and 4) naming a cluster using the features selected from the virtual documents. Experiments on pages (extracted from Yahoo! Categories) with less than 20 in-linked pages (obtained by Google) demonstrate very high accuracy of the proposed method.

There is one more justification for the more reliability of information originating from pages that point to the document than the features derived from the document text itself [5]. This work further distinguishes the different type of segments in each hyperlink pointing to a document, which is encoded with its anchor text, the headings structurally preceding it, and the text of the paragraph in which it occurs.

Many methods have been proposed for feature selection in text classification. In general, different linguistic components of a document can form different types of features. In addition to the use of a feature vector that consists of single tokens and does not consider the dependencies and the relative positions of different features, people also consider using *phrasal features* (consisting of more than one tokens), so as to take advantage of the token dependency and position information. While some experiments show that introducing some degree of term dependency in the Bayesian network method will achieve a higher accuracy comparing to the Naive Bayesian method with the independence assumption [14], [3], some other studies have been debating the opposite [11]. It is shown that a Naive Bayesian classifier with only noun phrases yielded much lower effectiveness than a classifier simply using bag-of-words. More sophisticated feature selection strategies include the approach using Information Gain measures, which are found fairly effective [20], [10]. It

is reported that term selection based on *document frequency* in the training set is simple but has similar performance to the Information Gain methods [20].

How to reduce the huge number of features in text classification methods is an important issue. Previous work can be divided into two categories: feature selection and re-parameterization. In feature selection, a subset of the most important features are selected from the entire feature space, to be used by the learning algorithm. Most previous work on classification has relied exclusively on this method. Re-parameterization is the process of constructing a new document representation by taking combinations and transformations of the original feature variables [16].

In the particular case of hierarchical classification, where categories are organized as a hierarchy, one approach to feature reduction has been proposed, which uses multiple Bayesian classifiers, each for one node of the hierarchy. The experiment shows that the hierarchical Bayesian classifier can restrict the large feature sets to be much smaller ones, thus providing higher accuracies than flat classifiers [9].

III. OUR METHOD

We propose an approach to select features by utilizing the citation information. The domain of our discussion is particularly of academic publications, in which the citation information of a publication has been thought of being a summarization of its content and an implication of its significance. Figure 1 shows the architecture of our proposed method.

A. Citation Contexts vs. Text Contents

Given two publications d_1 and d_2 , we say d_1 is a *citer* and d_2 is a *citee*, if d_1 cites d_2 . A *citation context* of a citee is a segment of the citer's text content, which contains at least one reference (or reference entry) to the citee. Such citation context can be a word sequence, a sentence, or even a paragraph. Our idea is to aggregate all the citation contexts for an identical publication so as to form an external representation (also called "virtual document" in previous work [6]) of that publication. The aggregation can be as straightforward as a concatenate of citation contexts. It can also be a syntactically valid and semantically meaningful summarization of the citation contexts using certain text summarization techniques [13].

Our strategy to testify the usefulness of citation information for text classification is to compare the performances of a classifier in three different cases, when it selects the features from three different sources: the text content, the citation contexts, and both of them. We choose to use Naive Bayes as our classifier and three simple feature selection strategies, as shown next.

B. Feature Selection Strategies

In our approach, we simply see a text as a bag of words, which is thus represented as a binary *feature vector*. The value of each item in the vector is either 1 if the feature appears in the observing document, or 0 otherwise.

Given text content, citation contexts, and both as the possible feature extraction sources, we use three different feature selection methods based on the consideration of *term frequency*: local-threshold based method, global-threshold based method, and most-frequent-term based method. A *local threshold* is used to select a token (or term) as a feature if its local frequency (times appearing in a document) exceeds a certain number (i.e., the threshold). Similarly, a *global threshold* is the one that is used to select a feature according to the term's global frequency. Differently from these two threshold based methods, most-frequent-term based method does not have a predefined threshold. Instead, it selects a certain number of features by choosing the terms that have the highest global frequency.

We note that in our system for experiments, the number of features to be selected will be specified by the classifier prior to the training. If any of the three feature selection strategies selects more features than the specified feature number, then some of the features will be filtered by picking candidate features at certain interval. For example, suppose that the feature selection algorithm outputs an array of 10000 terms as features, whereas the classifier only specifies to use 2000 features. Then, we pick the terms at the positions of $1 + i * (10000/2000)$, for $i \in [1..2000]$, as features. However, if the output terms are less than the specified feature number, we will then use all of them.

C. Naive Bayesian Classification

This section quickly reviews the basis of Naive Bayes. A Bayesian classifier is a Bayesian network applied to a classification domain. It contains a class variable C and a feature variable X_i for each of the features. Given an instance (i.e., a document to be classified) \mathbf{x} , which is an assignment of values of x_1, x_2, \dots, x_n to the feature variables, the Bayesian network allows us to compute the probability $P(C = c_k | \mathbf{X} = \mathbf{x})$ for each possible class c_k .

The simplest and earliest Bayesian classifier is the Naive Bayesian classifier, which is still widely employed nowadays. An important assumption that the Naive Bayesian classifier makes is that features are conditionally independent of one another, given the class variable. That means $P(\mathbf{X}|C) = \prod_i P(X_i|C)$. Consequently, the probability of a document \mathbf{X} belonging to a class c_k can be calculated by the following equation.

$$P(c_k | \mathbf{X}) = \frac{1}{Z} P(c_k) \prod_i P(X_i | c_k)$$

where Z is a scaling factor dependent only of the features.

IV. EXPERIMENTAL EVALUATION

A. Corpus

Our experiments use the Cora dataset, as was used in the Cora project at the University of Massachusetts. It contains a collection of academic publications in different areas of Computer Science, and covers basic publication attributes such as title, author, abstract, and references (including their

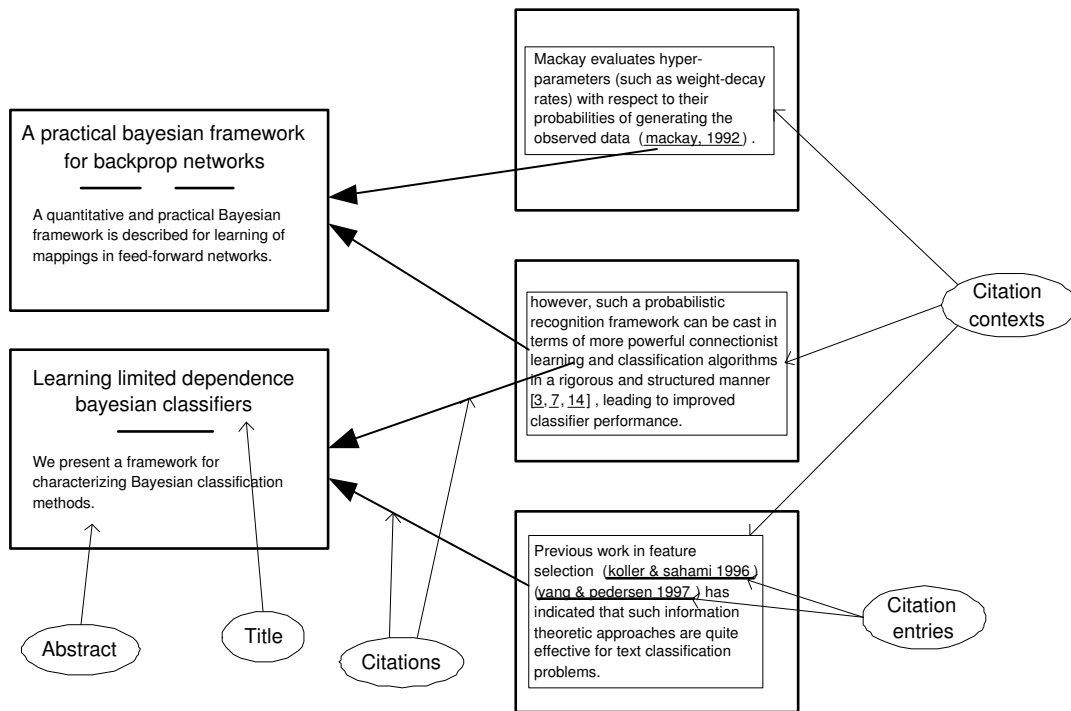


Fig. 1. Architecture.

entries and citation contexts). The category of each publication in the dataset is manually assigned, hence possibly causing subjective bias and inaccuracy.

The Cora dataset include totally 35,788 papers belonging to a hierarchy of classes, which has 10 classes at the top level and 68 leaf classes. Fortunately, besides the title and abstract, each publication in the dataset also has its references stored, from which we compute the citation contexts of the cited publication. However, there are 18,985 out of the entire collection being both citers and citees. Furthermore, there are only 12,605 papers out of these many papers having their category information available for use. Therefore, we finally have 12,605 papers ready for use, where few papers still have no abstracts. In our experiments, we use 9,000 papers as the training data and 3,000 as the testing objects.

We note that the 12,000 documents used for experiments still retain numerous typos and errors. In addition, the citation contexts are incomplete and sometimes inconsistent. Despite of all these noises, we ignore a more advanced preprocessing of data cleaning, for simplicity.

B. Experiment Results

We have performed a series of experiments on different combinations of the following factors of the feature selection process: (1) feature selection sources, which can be from the text content only, citation contexts only, and both, (2) feature selection strategies, including using a local threshold, a global threshold, or highest term frequency.

In the following presentation of our experimental results, the meaning of the following legends in Figure 2 denote different sources and strategies for feature selection: “content

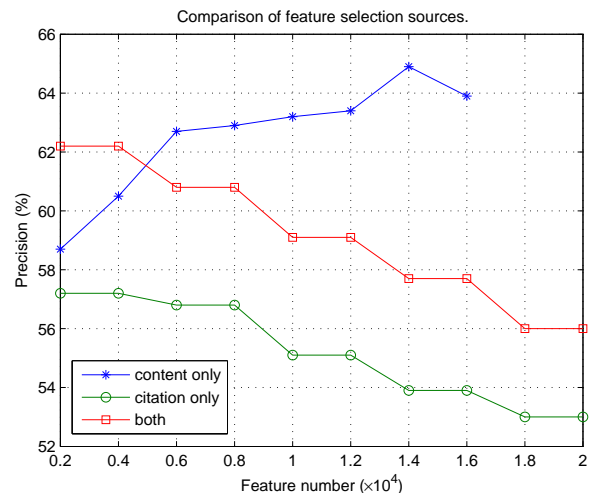


Fig. 2. Different feature selection sources.

only” standing for the feature selection from text contents only, “citation only” for that from citation contexts only, and “both” for that from both text and citation contexts.

Our first experiment is comparing the differences of using the three different sources for the feature selection: text contents, citation contexts, and both. The comparison is taken in terms of the precision (or accuracy) of the classification, which uses features selected from these three sources with a *global threshold* of 2. The experiment results are shown in Figure 2. We see that the highest precision 64.9% is led by the content-based feature selection at the point where 14,000

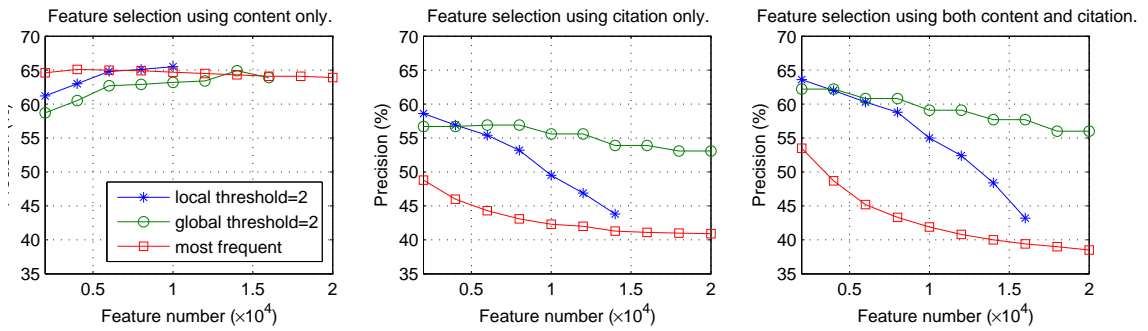


Fig. 3. Different feature selection strategies.

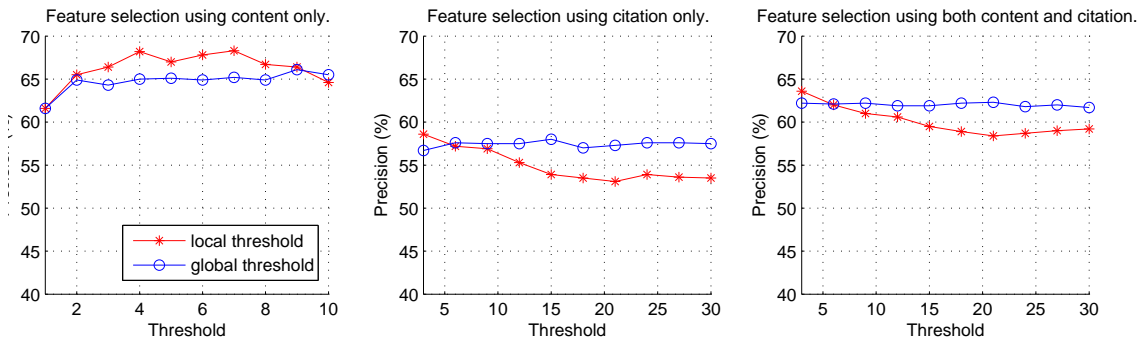


Fig. 4. Different thresholds.

features are used. We also notice that, while the precision of the content-based method increases with the feature number, the precision of the citation-based and both-based methods is non-increasing. It turns out that using citation contexts as a (sole or additional) source for feature selection can result in a considerable reduction of the feature number while keeping a reasonably high precision. Actually, the both-based method still has a precision of 62.2% even if using 2,000 features, which is higher than that (58.7%) of the content-based method using the same number of features.

In our second experiment, we try to find the differences of the three feature selection strategies, i.e., local-threshold based, global-threshold based, and most-frequent-term based. The classification precisions are evaluated using the features selected with local thresholds being 2 and global thresholds being 2, respectively. As shown in Figure 3, the experiment results are such as:

1) In the content-based method, local thresholds generally work better than global thresholds, with a proportionally higher precision (2.5 points higher). While the highest precision 65.5% is reached by using a local threshold at the point of 10,000 features, the precisions obtained by using the features selected from the most globally frequent terms are stably high on different feature numbers. In addition, the *micro-average precision* of using most frequent terms is 64.5%, generally higher than that of using local thresholds (63.9%).

2) Unsurprisingly, the precision of all three feature selection methods in citation-based and both-based approaches is nonincreasing with the feature numbers. It turns out that the

performance of using local thresholds is more sensitive to the feature numbers. Another interesting observation is that the precision of both-based approach is nearly proportionally higher than that of citation-based approach, in both cases of using local thresholds and of using global thresholds. This tells us an important implication: citation contexts of a text can act as a good representative, in the sense that their effects on classification performance will not change no matter whether the text content is considered or not. Finally, using most frequent terms for feature selection will result in many noises in both approaches, hence making the performance much worse.

Figure 4 shows the results of our third experiment with respect to the effects of different thresholds on the classification performances. In (both global and local) threshold-based feature selection methods, the number of features will decrease while the threshold increases. For each threshold, we evaluate the classification precision using an incremental number of features ranging from 2,000 to 20,000 at an interval of 2,000, and using all the features if they are less than 2,000. For comparison, we choose the highest precision for every global and local threshold. We find that the classification performance in all three cases of feature selection sources (content, citation, and both) are less affected by global thresholds than by local ones. In content-based case, the classification using local thresholds outperforms that using global thresholds in the almost entire range of tested thresholds. Whereas, we have the contrary results in the last two cases. We also find another way of reduce the feature number with the precision not dropping much, which

is achieved by selecting features from the content by means of a high local threshold. As a fact, the precision in this case can be as high as 66.4% by just using 502 features resulted from a local threshold of 9.

C. Discussion

The following gives a qualitative summarization of the above mentioned feature selection methods, based on the experimental results on them.

First, citation contexts, as external informative summarization of a text content, exhibit the following characteristics. On the one hand, they are defacto meaningful and concise representation of the text, thus providing a good source for the feature selection. This has been imprecisely proved by the coincidence of the both-based method with the citation-based method and by the higher precision of both-based method (compared to that of content-based method) when using a low threshold but few features. On the other hand, there still exist a large amount of noises (i.e., terms that are most unlikely to be features) in the citation contexts, which usually have a high local term frequency, as indicated by the third experiment.

Second, as we can observe in all experiments, the exploit of citation contexts for text classification exhibit a possible way to reduce the feature number, since the precision of the classification is non-decreasing with the feature number and is stable with different global thresholds. This means that even a few terms in the citation contexts over whatever a global threshold can lead to a reasonably high precision of the classification.

Besides the above observations about the use of citation contexts for feature selection, we also find out two other ways of feature number reduction. The first way is to use a high local threshold in the content-based method, as shown in Figure 4. The second way is to use a few number of most frequent terms from the text content. The latter has an additional advantage that the precision is stable with different feature numbers.

The last point that we want to make is that the overall classification performance of the both-based method is lower than that of the content-based method. The reasons may include: the moderate quality of the corpus, the much naive feature selection algorithm based on a threshold and a random picking algorithm, and lacking of a sophisticated analysis on the citation contexts.

V. CONCLUSIONS AND FUTURE WORK

Feature selection has been an important issue in the text classification using statistical methods, such as Naive Bayes. The most difficult problem in feature selection is how to select as few as features that has greater discriminative and descriptive power. In this paper, we propose an feature selection approach, by exploiting external information that summarizes the text to be classified. We particularly study the use of *citation contexts* of academic publications for their

categorization using the Naive Bayesian method. Our experiment results allow us to make some important observations as follows:

First, citation contexts are indeed meaningful and descriptive representatives of a text, thus providing a good source for feature selection. This enhances the results in previous work on citation analysis. Second, we find several unsophisticated ways to reduce the feature number while keeping a high classification precision, such as: using a low local threshold in the both-based method, using a high local threshold in the content-based method, and using few most frequent terms extracted from the text content.

In future, in order to better utilize the citation information for text classification, we have to uncover the reasons behind the fact that the overall classification performance of the both-based method is lower than that of the content-based method. However, before going further in this way, we would first perform more experiments on the citation (or both) based feature selection strategy using fewer features than 2,000. This is because of the observation on the first experiment result that the citation-based (or both-based) classification tends to have a higher accuracy (than content-based classification) when the number of features is smaller. We will also take a close look at the experiment results in this paper, to try to find a way combining the three unsophisticated feature selection methods into a more effective feature selection method, which will just use a few features.

REFERENCES

- [1] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107-117, 1998.
- [2] P. Calado, M. Cristo, E. S. de Moura, N. Ziviani, B. A. Ribeiro-Neto, and M. A. Gonçalves. Combining Link-based and Content-based Methods for Web Document Classification. In *CIKM 2003*, pages 394-401.
- [3] S. T. Dumais, J. C. Platt, D. Hecherman, and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. In *CIKM 1998*, pages 148-155.
- [4] L. Egghe and R. Rousseau. Co-citation, Bibliographic Coupling and a Characterization of Lattice Citation Networks. *Scientometrics*, 55(3):349-361, 2002.
- [5] J. Fürnkranz. Exploiting Structural Information for Text Classification on the WWW. In *Third International Symposium on Advances in Intelligent Data Analysis (IDA)*, pages 487-498, 1999.
- [6] E. J. Glover, K. Tsioutsoulouklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web Structure for Classifying and Describing Web Pages. In *WWW 2002*, pages 562-569.
- [7] T. H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784-796, 2003.
- [8] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998)*, pages 668-677, 1998.
- [9] D. Koller and M. Sahami. Hierarchically Classifying Documents Using Very Few Words. In *ICML 1997*, pages 170-178.
- [10] D. Koller and M. Sahami. Toward Optimal Feature Selection. In *ICML 1996*, pages 284-292.
- [11] D. D. Lewis. Feature Selection and Feature Extraction for Text. In *Proceedings of Speech and Natural Language Workshop*, 1992.
- [12] H. Nanba, N. Kando, and M. Okumura. Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. In *Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop on Classification for User Support and Learning*, pages 117-134, 2000.

- [13] H. Nanba and M. Okumura. Towards Multi-paper Summarization Using Reference Information. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI 1999)*, pages 926–931, 1999.
- [14] M. Sahami. Learning Limited Dependence Bayesian Classifiers. In *KDD 1996*, pages 335–338.
- [15] G. Salton. Associative Document Retrieval Techniques Using Bibliographic Information. *Journal of the ACM*, 10(4):440–457, 1963.
- [16] H. Schütze, D. A. Hull, and J. O. Pedersen. A Comparison of Classifiers and Document Representations for the Routing Problem. In *SIGIR 1995*, pages 284–292.
- [17] W. Tao and W. Zuo. Query-Sensitive Self-Adaptable Web Page Ranking Algorithm. In *Proceedings of the Second International Conference on Machine Learning and Cybernetics (ICMLC 2003)*, pages 413–418, 2003.
- [18] S. Teufel and M. Moens. Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, 1999.
- [19] L. Vaughan and D. Shaw. Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology (JASIST)*, 54(14):1313–1322, 2003.
- [20] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML 1997*, pages 412–420.