An Extended Document Frequency Metric for Feature Selection in Text Categorization

Yan Xu, Bin Wang, JinTao Li, and Hongfang Jing

Institute of Computing Technology,Chinese Academy of Sciences
No.6 Kexueyuan South Road, Zhongguancun,Haidian District, Beijing,China
{xuyan,wangbin,jtli,jinghongfang}@ict.ac.cn

Abstract. Feature selection plays an important role in text categorization. Many sophisticated feature selection methods such as Information Gain (IG), Mutual Information (MI) and $\chi 2$ statistic measure (CHI) have been proposed. However, when compared to these above methods, a very simple technique called Document Frequency thresholding (DF) has shown to be one of the best methods either on Chinese or English text data. A problem is that DF method is usually considered as an empirical approach and it does not consider Term Frequency (TF) factor. In this paper, we put forward an extended DF method called TFDF which combines the Term Frequency (TF) factor. Experimental results on Reuters-21578 and OHSUMED corpora show that TFDF performs much better than the original DF method.

Keywords: Rough Set, Text Categorization, Feature Selection, Document Frequency.

1 Introduction

Text categorization is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data.

A major difficulty of text categorization is the high dimensionality of the original feature space. Consequently, feature selection-reducing the original feature space,is seriously projected and carefully investigated.

In recent years, a growing number of statistical classification methods and machine learning techniques have been applied in this field. Many feature selection methods such as document frequency thresholding, information gain measure, mutual information measure, $\chi 2$ statistic measure, and term strength measure have been widely used.

DF thresholding, almost the simplest method with the lowest cost in computation, has shown to behave comparably well when compared to more sophisticated statistical measures [13], it can be reliably used instead of IG or CHI while the computation of these measures are more expensive. Especially, experiments show that DF has better performance in Chinese text categorization [1][11] than IG, MI and CHI. In one

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 71–82, 2008.

word, DF, though very simple, is one of the best feature selection methods either for Chinese or English text categorization.

Due to its simplicity and effectiveness, DF is adopted in more and more experiments[7][4][2][6]. However, this method is only based on an empirical assumption that rare terms are noninformative for category prediction. In addition, like most feature selection methods, DF does not consider the Term Frequency (TF) factor, which is considered to be a very important factor for feature selection[12].

Rough Set theory, which is a very useful tool to describe vague and uncertain information, is used in this paper to give a theoretical interpretation of DF method. In Rough Set theory, knowledge is considered as an ability to partition objects. We then quantify the ability of classify objects, and call the amount of this ability as knowledge quantity. We use the knowledge quantity of the terms to rank them, and then put forward an extended DF method which considers the term frequency factor. Experiments show the improved method has notable improvement in the performances than the original DF.

2 Document Frequency Thresholding and Rough Set Theory Introduction

2.1 Document Frequency Thresholding

A term's document frequency is the number of documents in which the term occurs in the whole collection. DF thresholding is computing the document frequency for each unique term in the training corpus and then removing the terms whose document frequency are less than some predetermined threshold. That is to say, only the terms that occur many times are retained. DF thresholding is the simplest technique for vocabulary reduction. It can easily scale to very large corpora with a computational complexity approximately linear in the number of training documents.

At the same time, DF is based on a basic assumption that rare terms are noninformative for category prediction. So it is usually considered an empirical approach to improve efficiency. Obviously, the above assumption contradicts a principle of information retrieval (IR), where the terms with less document frequency are the most informative ones [9].

2.2 Basic Concepts of Rough Set Theory

Rough set theory, introduced by Zdzislaw Pawlak in 1982 [5][8], is a mathematical tool to deal with vagueness and uncertainty. At present it is widely applied in many fields, such as machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, pattern recognition, etc. In this section, we introduce some basic concepts of rough set theory which used in this paper.

Given two sets U and A, where $U = \{x_1, ..., x_n\}$ is a nonempty finite set of objects called the universe, and $A = \{a_1, \ldots, a_k\}$ is a nonempty finite set of attributes, the

attributes in A is further classified into two disjoint subsets, condition attribute set C and decision attribute set D, $A=C \cup D$ and $C \cap D = \Phi$. Each attribute $a \in A$, V is the domain of values of A, V_a is the set of values of a, defining an information function f_a, : $U \rightarrow V_a$, we call 4-tuple <U,A,V, f > as an information system. a(x) denotes the value of attribute a for object x .

Any subset $B \subseteq A$ determines a binary relation Ind(B) on U, called indiscemibility relation:

$$Ind(B)=\{ (x,y) \in U \times U \mid \forall a \in B , a(x) = a(y) \}$$

The family of all equivalence classes of Ind(B), namely the partition determined by B, will be denoted by U/B . If (x , y) $\in$ Ind(B), we will call that x and y are B-indiscernible .Equivalence classes of the relation Ind(B) are referred to as B - elementary sets.

3 A Rough Set Interpretation of Document Frequency Thresholding

Given a 4-tuple <U,A,V, f > information system for text categorization, where $U=\{D_1, ..., D_n\}$ is a set of documents, $A = \{ t_1,\ldots, t_k \}$ is a set of features (terms), V is the domain of values of t_i ($1 \le i \le k$), $V=\{0,1\}$, An information function f, $U \rightarrow V$, can be defined as:

$$f(D_i) = \begin{cases} 0, & t \ dosen't \ occurs \ in \ D_i \\ 1, & t \ occurs \ in \ D_i \end{cases}$$

An example of such an information table is given in Table 1. Rows of Table 1, labeled with D_1, D_2, …D_6, are documents, the features are T_1, T_2, T_3 and T_4.

3.1 The Ability to Discern Objects

The important concept in rough set theory is indiscernibility relation. For example, in Table 1, (D_1, D_2) is T_1-indiscernible, (D_1, D_3) is not T_1-indiscernible.

Table 1. An information table: terms divide the set of documents into two equivalence classes

	T_1	T_2	T_3	T_4
D_1	0	0	1	1
D_2	0	1	0	1
D_3	1	0	1	1
D_4	0	0	1	1
D_5	0	0	0	1
D_6	0	1	0	1

In Table 1, T_1 only occurs in D_3, so T_1 divides $\{D_1, D_2, \ldots D_6\}$ into two equivalence classes $\{D_1, D_2, D_4, D_5, D_6\}$ and $\{D_3\}$. That is to say, T_1 can discern D_3 from D_1, D_2, D_4, D_5, D_6. Similarly, T_2 can discern D_2, D_4 from D_1, D_3, D_5, D_6. T_3 can discern D_1, D_3, D_4 from D_2, D_5, D_6. T_4 can not discern each document from the other, because T_4 divides $\{D_1, D_2, \ldots D_6\}$ into only one equivalence class. Now we quantify the ability of discerning objects for a feature T_i or a set of features P, we call the amount of the ability of discerning objects as *knowledge quantity*.

3.2 Knowledge Quantity

This section will be discussed on information table (Let decision feature set $D = \Phi$).

Definition 1. The object domain set U is divided into m equivalence classes by the set P (some features in information table), the number of elements in each equivalence class is: $n_1, n_2, \ldots, n_m$, let $W_{U,P}$ denotes the knowledge quantity of P, $W_{U,P}=W(n_1, n_2, \ldots, n_m)$, and it satisfies the following conditions:

1) $W(1,1) = 1$
2) if $m = 1$ then $W(n_1)=W(n) = 0$
3) $W(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m) = W(n_1, \ldots, n_j, \ldots, n_i, \ldots, n_m)$
4) $W(n_1,n_2, \ldots, n_m) = W(n_1,n_2+\ldots+n_m) + W(n_2,\ldots,n_m)$
5) $W(n_1,n_2+n_3) = W(n_1,n_2) + W(n_1,n_3)$

Conclusion 1. If the domain U is divided into m equivalence classes by some feature set P, and the element number of each equivalence class is $n_1,n_2, \ldots n_m$, then the knowledge quantity of P is: $W(n_1,n_2,\ldots,n_m)= \sum_{1 \le i < j \le m} n_i \times n_j$.

3.3 Interpretation of Document Frequency Thresholding

In Table 1, T_1 only occurs in D_3, T_1 divides $\{D_1, D_2, \ldots D_6\}$ into two equivalence classes $\{D_1, D_2, D_4, D_5, D_6\}$ and $\{D_3\}$, the number of each equivalence classes is $n_1=5$, $n_2=1$. According to Conclusion 1, the ability of discern $\{D_1, D_2, \ldots, D_6\}$ for T_1 (the knowledge quantity of T_1) is: $W_{U,T_1} = \sum_{1 \le i < j \le 2} n_i \times n_j = 5 \times 1 = 5$

Let U denote a set of all documents in the corpus, n denotes the number of documents in U, m demotes the number of documents in which term t occurs, the knowledge quantity of t is defined to be:

$$W_{U,t} = m(n-m) \tag{1}$$

$\because$ m=DF

$\therefore$ When $m \le n/2$, DF $\propto W_{U,t}$

After stop words removal, stemming, and converting to lower case, almost all term's DF value is less than n/2.

We compute the knowledge quantity for each unique term by (1) in the training corpus and remove those terms whose knowledge quantity are less than some predetermined threshold, this is our Rough set-based feature selection method which do not consider term frequency information(RS method). Feature selected by DF is the same as selected by this RS method. This is an interpretation of DF method.

4 An Extended DF Method Based on Rough Set

DF method does not consider the term frequency factor, however, a term with high frequency in a document should be more informative than the term that occurs only once. So, we think that terms divide the set of documents into not only two equivalence classes, but more than two equivalence classes, a term occurs in a document at least twice should be different from once occurs in the document, so there are 3 equivalence classes in our method.

Given a 4-tuple <U,A,V, f > information system, U={D_1, ..., D_n} is a set of documents, A = { t_1,…, t_k } is a set of features (terms), V is the domain of values of t_i (1≤i≤k), V={0,1,2}, defining an information function f, U→V:

$$f(D_i) = \begin{cases} 0, & t \text{ dosen't occurs in } D_i \\ 1, & t \text{ once occurs in } D_i; \ 2, \ t \text{ occurs in } D_i \text{ at least twice} \end{cases}$$

An example of such an information system is given in Table 2.

Table 2. An information table: terms divide the set of documents into three equivalence classes

	T_1	T_2
D_1	1	1
D_2	0	0
D_3	0	0
D_4	0	0
D_5	1	2
D_6	0	0

In Table 2, term T_1 occurs once both in D_1 and D_5, T_2 occurs once in D_1 but occurs more than once in D_5, the document frequency of term T_1 and T_2 is the same, but

$$W_{U,T_1} = \sum_{1 \le i < j \le 2} n_i \times n_j = 2 \times 4 = 8$$

$$W_{U,T_2} = \sum_{1 \le i < j \le 3} n_i \times n_j = (1 \times 1 + 1 \times 4 + 1 \times 4) = 9,$$

$$W_{U,T_2} > W_{U,T_1}$$

Let n denotes the number of documents in the corpus, term t divides the documents into 3 equivalence classes, and the number of elements in each equivalence class is: n_1, n_2, n_3. n_1 denotes the number of documents which t does not occurs, n_2 denotes the

number of documents which t occurs only once, n_3 denotes the number of documents which t occurs at least twice. The knowledge quantity of t is defined as:

$$W_{U,t} = \sum_{1 \le i < j \le 3} n_i \times n_j \tag{2}$$

In order to emphasize the importance of multiple occurences of a term, equation (2) can be changed to:

$$\mathrm{TFDF(t)} = (n_1 \times n_2 + c(n_1 \times n_3 + n_2 \times n_3)) \tag{3}$$

Here c is a constant parameter($c \ge 1$). As the value of c increases, we give more weight for multiple occurrences of a term.

Given a training corpus, we compute the TFDF(t) by (3) for all terms and rank them, then remove those terms which are in an inferior position from the feature space, this is our feature selection method based on rough set theory, we call it term frequency-based document frequency(TFDF).

5 Experimental Results

Our objective is to compare the original DF with the TFDF method. A number of statistical classification and machine learning techniques have been applied to text categorization, we use two different classifiers, k-nearest-neighbor classifier (kNN) and Naïve Bayes classifier. We use kNN, which is one of the top-performing classifiers, evaluations [14] have shown that it outperforms nearly all the other systems, and we selected Naïve Bayes because it is also one of the most efficient and effective inductive learning algorithms for classification [16]. According to [15], micro-averaging precision was widely used in Cross-Method comparisons, here we adopt it to evaluate the performance of different feature selection methods.

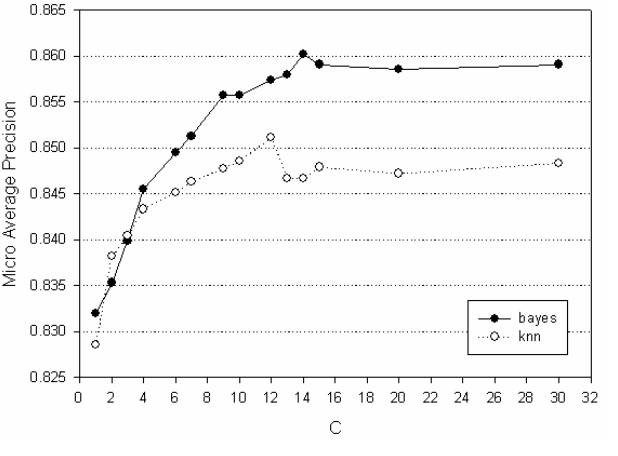
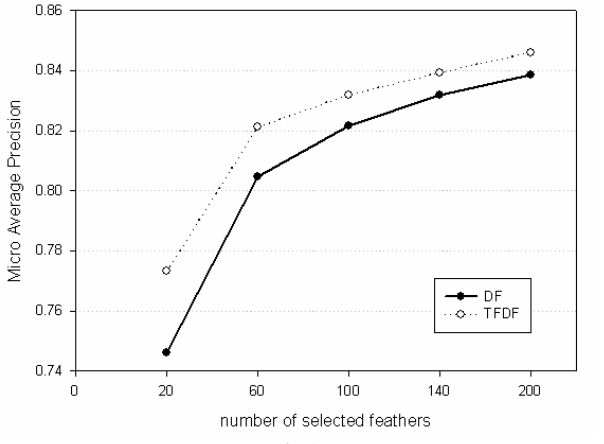


Fig. 1. Average precision of KNN and Naïve Bayes vary with the parameter c in Reuters (using TFDF method)

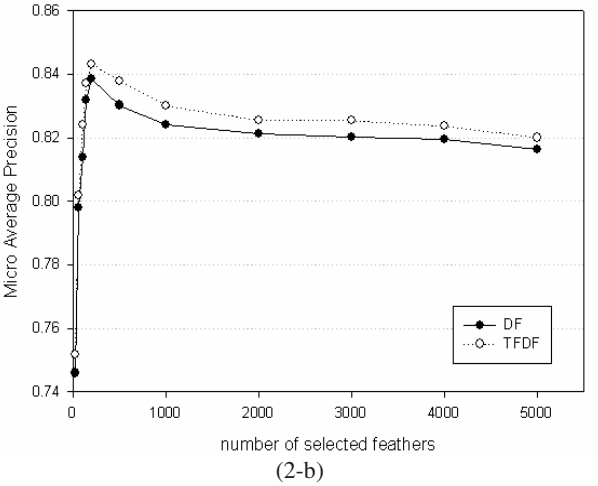
5.1 Data Collections

Two corpora are used in our experiments: Reuters-21578 collection[17] and the OHSUMED collection[19].

The Reuters-21578 collection is the original Reuters-22173 with 595 documents which are exact duplicates removed, and has become a new benchmark lately in text categorization evaluations. There are 21578 documents in the full collection, less than half of the documents have human assigned topic labels. In our experiment, we only consider those documents that had just one topic, and the topics that have at least 5 documents. The training set has 5273 documents, the testing set has 1767 documents. The vocabulary number is 13961 words after stop words removal, stemming, and converting to lower case.



(2-a)



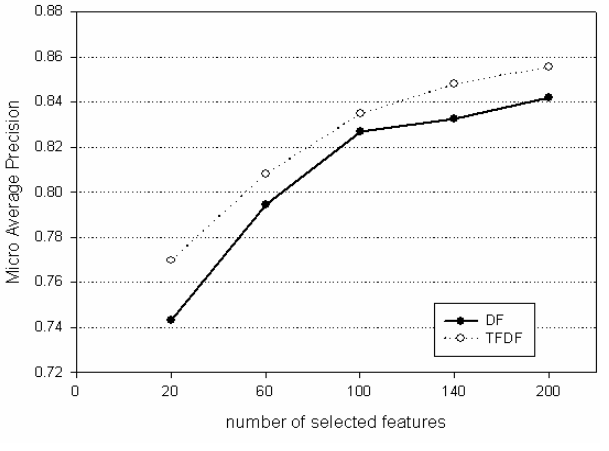
(2-b)

Fig. 2. Average precision of KNN vs. DF and TFDF number of selected features in Reuters

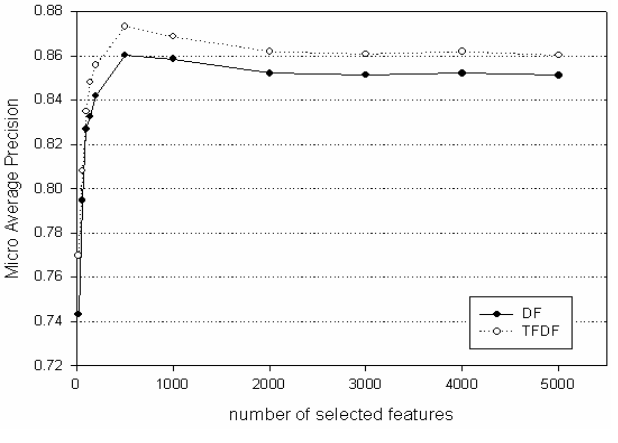
OHSUMED is a bibliographical document collection. The documents were manually indexed using subject categories in the National Library of Medicine. There are about 1800 categories defined in MeSH, and 14321 categories present in the OHSUMED document collection. We used a subset of this document collection. 7445 documents as a training set and the 3729 documents as the test set in this study. There are 11465 unique terms in the training set and 10 categories in this document collection.

5.2 Experimental Setting

Before evaluating the feature selection methods, we use the same selected feature number in both DF method and the TFDF method for the experiment. Weka[18] is used as our experimental platform.



(3-a)



(3-b)

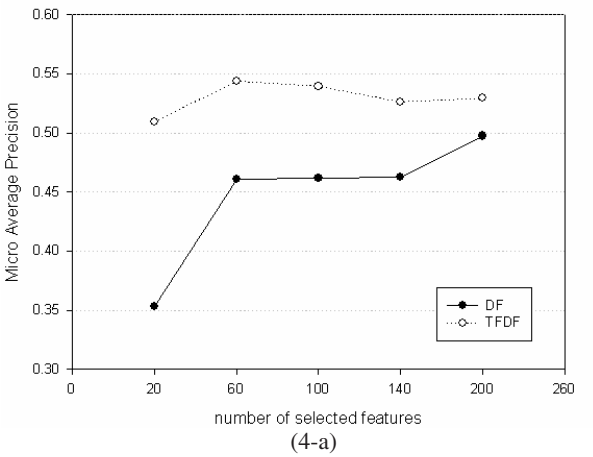
Fig. 3. Average precision of Naïve Bayes vs. DF and RS number of selected features in Reuters

5.3 Results

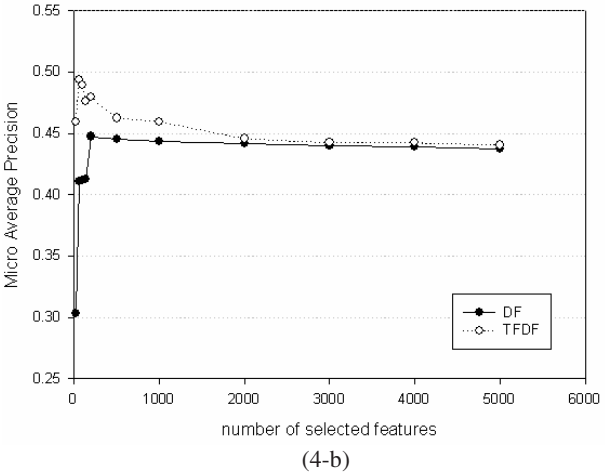
Figure 1 shows that the Average precision of KNN and Naïve Bayes varying with the parameter c in Reuters (in equation (3), using TFDF method) at a fixed number of selected features, here, the fixed number of selected features is 200. We can notice that when c≤12, as c increases, the Average precision increases accordingly.

Figure 2 and Figure 3 exhibit the performance curves of kNN and Naïve Bayes on Reauters-21578 after feature selection DF and TFDF(c=10). We can note from figure 2 and figure 3 that TFDF outperform DF methods, specially, on extremely aggressive reduction, it is notable that TFDF prevalently outperform DF((2-a),(3-a)).

Figure 4 and Figure 5 exhibit the performance curves of kNN and Naïve Bayes on OHSUMED after feature selection DF and TFDF(c=10). We can also note from figure 2 and figure 3 that TFDF outperform DF methods, specially, on extremely aggressive reduction, it is notable that TFDF prevalently outperform DF((4-a),(5-a)).

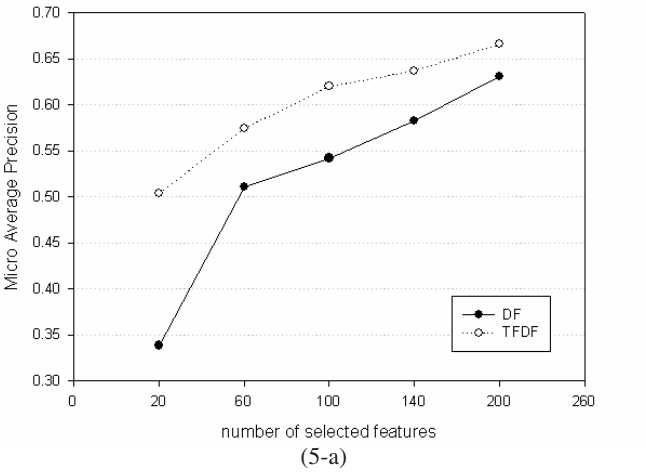


(4-a)

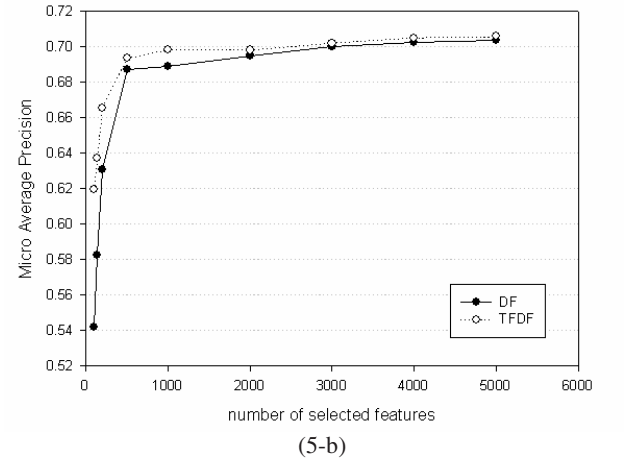


(4-b)

Fig. 4. Average precision of KNN vs. DF and TFDF number of selected features on OHSUMED



(5-a)



(5-b)

Fig. 5. Average precision of Naïve Bayes vs. DF and TFDF number of selected features on OHSUMED

6 Conclusion

Feature selection plays an important role in text categorization. DF thresholding, almost the simplest method with the lowest cost in computation, has shown to behave well when compared to more sophisticated statistical measures, However, DF method is usually considered as an empirical approach and does not have a good theoretic interpretation, and it does not consider Term Frequency (TF) factor, in this paper: we put forward an extended DF method called TFDF which combines the Term Frequency (TF) factor. Experiments on Reuters-21578 collection and OHSUMED

collection show that TFDF perform much better than the original DF method, specially, on extremely aggressive reduction, it is notable that TFDF prevalently outperform DF. The experiments also show that Term Frequency factor is important for feature selection.

Many other feature selection methods such as information gain measure, mutual information measure, $\chi 2$ statistic measure, and term strength measure have been widely used in text categorization, but none of them consider Term Frequency (TF) factor, In the future research we will investigate to use TF in these feature selection methods.

Acknowledgments. This work is supported by the National Natural Science Fundamental Research Project of China (60473002, 60603094), the National 973 Project of China (2004CB318109)and the National Natural Science Fundamental Research Project of Beijing (4051004).

References

1. Liu-ling, D., He-yan, H., Zhao-xiong, C.: A comparative Study on Feature Selection in Chinese Text Categorization. Journal of Chinese Information Processing 18(1), 26–32 (2005)
2. Dumais, S.T., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Proceedings of CIKM 1998, pp. 148–155 (1998)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
4. Itner, D.J., Lewis, D.D.: Text categorization of low quality images. In: Proceedings of SDAIR 1995, pp. 301–315 (1995)
5. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In: A New Trend in Decision-Making, pp. 3–98. Springer, Singapore (1999)
6. Li, Y.H., Jain, A.K.: Classification of text documents. Comput. J. 41(8), 537–546 (1998)
7. Maron, M.: Automatic indexing: an experimental inquiry. J. Assoc. Comput. Mach. 8(3), 404–417 (1961)
8. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Science 11(5), 341–356 (1982)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inform. Process. Man 24(5), 513–523 (1988)
10. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
11. Songwei, S., Shicong, F., Xiaoming, L.: A Comparative Study on Several Typical Feature Selection Methods for Chinese Web Page Categorization. Journal of the Computer Engineering and Application 39(22), 146–148 (2003)
12. Yang, S.M., Wu, X.-B., Deng, Z.-H., Zhang, M., Yang, D.-Q.: Modification of Feature Selection Methods Using Relative Term Frequency. In: Proceedings of ICMLC 2002, pp. 1432–1436 (2002)
13. Yang, Y., Pedersen, J.O.: Comparative Study on Feature Selection in Text Categorization. In: Proceedings of ICML 1997, pp. 412–420 (1997)

14. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: SIGIR 1999, pp. 42–49 (1999)
15. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1(1/2), 67–88 (1999)
16. Zhang, H.: The optimality of naive Bayes. In: The 17th International FLAIRS conference, Miami Beach, May 17-19 (2004)
17. Reuters 21578, http://www.daviddlewis.com/resources/testcollections/reuters21578/
18. Weka, http://www.cs.waikato.ac.nz/ml/weka/
19. OHSUMED, http://www.cs.umn.edu/%CB%9Chan/data/tmdata.tar.gz