New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents

Li-Wei Lee and Shyi-Ming Chen

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan, R.O.C.
smchen@et.ntust.edu.tw

Abstract. In this paper, we present a new feature selection method based on document frequencies and statistical values. We also present a new similarity measure to calculate the degree of similarity between documents. Based on the proposed feature selection method and the proposed similarity measure between documents, we present three methods for dealing with the Reuters-21578 top 10 categories text categorization. The proposed methods get higher performance for dealing with the Reuters-21578 top 10 categories text categorization than that of the method presented in [4].

1 Introduction

Text categorization is a task of classifying text documents into a predefined number of categories based on classification patterns [3], [22]. The terms appearing in documents are treated as features. One major difficulty in text categorization is the large dimension of the feature space. Therefore, we hope to reduce the dimension of the feature space to get a higher performance for dealing with the text categorization problem. One method for text categorization is based on the feature selection method [3], [4], [21]. Some results from the previous researches show that the semantic feature selection approach affects the performance of text categorization [3], [21]. Some feature selection methods have been presented to deal with a wide range of text categorization tasks, such as Chi-Square test [1], [2], [6], [20], Information Gain (IG) [1], [9], [10], [12], [15], and Mutual Information (MI) [2], [5], [8], [12], [13].

In this paper, we present a new feature selection method based on document frequencies and statistical values to select useful features. We also present a new similarity measure between documents. Based on the proposed feature selection method and the proposed similarity measure between documents, we present three methods for dealing with the Reuters-26578 top 10 categories text categorization. The proposed methods can get higher performance for dealing with the Reuters-26578 top 10 categories text categorization than that of the method presented in [4].

The rest of this paper is organized as follows. In Section 2, we briefly review the previous research for text categorization. In Section 3, we present a new feature selection

M. Ali and R. Dapoigny (Eds.): IEA/AIE 2006, LNAI 4031, pp. 1280 – 1289, 2006.

method based on document frequencies and statistical values for text categorization. In Section 4, we present a new similarity measure between documents. Based on the proposed feature selection method and the proposed similarity measure between documents, we present three methods to deal with the Reuters-26578 top 10 categories text categorization. In Section 5, we show the experimental results. The conclusions are discussed in Section 6.

2 Preliminaries

In the vector space model [18], documents are usually represented by feature vectors of terms. The task of preprocessing consists of transforming capital letters into lowercased letters and removing stop words (such as "a", "an", "the", etc.), where words are stemmed by applying the Porter algorithm [17]. The acquired document-term matrix is then transformed into TF-IDF (Term Frequency-Inverse Document Frequency) weights which are normalized by document lengths [19]. After the feature selection process, the dimension of the feature space is reduced and useful features are obtained. Therefore, the feature selection process is a very important task, and it can affect the performance of text categorization.

Assume that F consists of n features $f_1, f_2, \ldots, f_n$ and assume that S consists of m features $s_1, s_2, \ldots, s_m$, where S is a subset of F. The goal of the feature selection process is to choose an optimal subset S of F for text categorization. There are many statistic measures for dealing with the task of feature selection, e.g., Chi-Square [1], [2], [6], [20], Information Gain [1], [9], [10], [12], [15], and Mutual Information [2], [5], [8], [12], [13]. Among these measures, the mutual information measure is the most commonly used measure. It also has a better performance for dealing with the task of feature selection. In the following, we briefly review some feature selection measures, shown as follows:

(1) Chi-Square test [2]: Fix a term t, let the class labels be 0 and 1. Let $k_{i,0}$ denote the number of documents in class i not containing term t and let $k_{i,1}$ denote the number of documents in class i containing term t. This gives us a 2×2 contingency matrix:

		I_t	
		0	1
C	0	k_{00}	k_{01}
	1	k_{10}	k_{11}

where C and I_t denote Boolean random variables and $k_{\ell m}$ denotes the number of observations, where $C \in \{0,1\}$ and $I_t \in \{0,1\}$. Let $n = k_{00} + k_{01} + k_{10} + k_{11}$. We can estimate the marginal distribution as follows:

$$\Pr(C = 0) = (k_{00} + k_{01})/n,$$
$$\Pr(C = 1) = (k_{10} + k_{11})/n,$$
$$\Pr(I_t = 0) = (k_{00} + k_{10})/n,$$
$$\Pr(I_t = 1) = (k_{01} + k_{11})/n.$$

The χ^2 test is shown as follows:

$$\chi^2 = \sum_{\ell,m} \frac{(k_{\ell m} - n \Pr(C = \ell) \Pr(I_t = m))^2}{n \Pr(c = \ell) \Pr(I_t = m)} = \frac{n(k_{11}k_{00} - k_{10}k_{01})^2}{(k_{11} + k_{10})(k_{01} + k_{00})(k_{11} + k_{01})(k_{10} + k_{00})}. \quad (1)$$

The larger the value of χ^2, the lower is our belief that the independence assumption is upheld by the observed data. In [2], Chakrabarti pointed out that for feature selection, it is adequate to sort terms in decreasing order of their χ^2 values, train several classifiers with a varying number of features, and stopping at the point of maximum accuracy (see [2], pp. 139). The larger the value of χ^2, the higher the priority to choose term *t*. For more details, please refer to [2].

(2) Information Gain Measure [23], [24]: For the binary document model and two classes (the same as the case of the χ^2 test), the Information Gain (IG) of term t with respect to the two classes can be written as follows:

$$IG(t) = -\sum_{i=0}^{1} P(c_i) \log P(c_i) - (-P(t) \sum_{i=0}^{1} P(c_i \mid t) \log P(c_i \mid t) - P(\bar{t}) \sum_{i=0}^{1} P(c_i \mid \bar{t}) \log P(c_i \mid \bar{t})), \quad (2)$$

where $P(c_0) = \frac{k_{00} + k_{01}}{k_{00} + k_{01} + k_{10} + k_{11}}$, $P(c_1) = \frac{k_{10} + k_{11}}{k_{00} + k_{01} + k_{10} + k_{11}}$,

$P(t) = \frac{k_{00} + k_{10}}{k_{00} + k_{01} + k_{10} + k_{11}}$, $P(\bar{t}) = \frac{k_{01} + k_{11}}{k_{00} + k_{01} + k_{10} + k_{11}}$, $P(c_0 \mid t) = \frac{k_{00}}{k_{00} + k_{10}}$,

$P(c_1 \mid t) = \frac{k_{10}}{k_{00} + k_{10}}$, $P(c_0 \mid \bar{t}) = \frac{k_{01}}{k_{01} + k_{11}}$, and $P(c_1 \mid \bar{t}) = \frac{k_{11}}{k_{01} + k_{11}}$.

The larger the value of $IG(t)$, the higher the priority to choose term *t*. For more details, please refer to [23] and [24].

(3) Mutual Information Measure [2]: For the binary document model and two classes (the same as the case of the χ^2 test), the Mutual Information (MI) of term t with respect to the two classes can be written as follows:

$$MI(I_t, C) = \sum_{\ell,m \in \{0,1\}} \frac{k_{\ell,m}}{n} \log \frac{k_{\ell,m} / n}{(k_{\ell,0} + k_{\ell,1})(k_{0,m} + k_{1,m}) / n^2}, \quad (3)$$

where $n = k_{00} + k_{01} + k_{10} + k_{11}$. The larger the value of $MI(I_t, C)$, the higher the priority to choose term *t*. For more details, please refer to [2].

3 A New Feature Selection Method Based on Statistical Values and Document Frequencies

In this section, we present a new feature selection method based on statistical values and document frequencies. Let X and Y be two different classes of documents. The mean values $\mu_{X,t}$ and $\mu_{Y,t}$ of X and Y are $1/|X|(\Sigma_X x_t)$ and $1/|Y|(\Sigma_Y y_t)$, respectively, where x_t denotes the TFIDF of term t in class X and y_t denotes the TFIDF of term t in class Y. Furthermore, the variances σ_X and σ_Y of X and Y are $1/|X|\Sigma_X(x_t-\mu_{X,t})^2$ and $1/|Y|\Sigma_Y(y_t-\mu_{Y,t})^2$, respectively. Let |X| denote the number of documents in the class X and let |Y| denote the number of documents in the class Y. Here, we consider the effect of document frequencies and variances for feature selection. Let $DF(x_t)$ denote the document frequencies of term x_t in the X class and let $DF(y_t)$ denote the document frequency of term y_t in the Y class. The proposed feature selection method is as follows:

$$S(t)=\frac{(DF(x_t)/|X|-DF(y_t)/|Y|)^2}{(1/|X|)\Sigma_X(x_t-\mu_{X,t})^2+(1/|Y|)\Sigma_Y(y_t-\mu_{Y,t})^2}. \tag{4}$$

The larger the value of S(t), the higher the priority to choose term t.

4 New Methods for Text Classification Based on the Proposed Similarity Measure and the k-NN Approach

Many learning-based approaches have been presented to deal with the task of text categorization, e.g., the k-NN approach [7], [22], [24], support vector machines [5], [14], [24], Naïve Bayes approaches [5], [10], [24], and neural networks [16], [24]. In this paper, we present three classification methods based on the k-NN approach [7], [22], [24] to classify the Reuters-21578 top 10 categories data set.

The k-NN classifier uses the k-nearest training documents with respect to a testing document to calculate the likelihood of categories. The document-document similarity measure used in the k-NN classifier is the most important part for text categorization. Most previous k-NN classifiers use the cosine similarity measure in the vector space model. The cosine similarity measure $\cos(\vec{a},\vec{b})$, for measuring the degree of similarity between documents a and b is as follows:

$$\cos(\vec{a},\vec{b})=\frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\cdot\|\vec{b}\|}, \tag{5}$$

where $\cos(\vec{a},\vec{b})\in[0,1]$, $\vec{a}$ and $\vec{b}$ denote the vector representation of the documents a and b, respectively. The larger the value of $\cos(\vec{a},\vec{b})$, the higher the similarity between the documents a and b.

The term weight w_{ij} calculated by TFIDF normalized by document lengths is the most commonly-used method [19] for the cosine similarity measure in the vector space model, where

$$w_{ij} = \frac{TFIDF(t_i, d_j)}{\sqrt{\sum_{k=1}^{|T|} (TFIDF(t_k, d_j))^2}}, \tag{6}$$

w_{ij} denotes the weight of term i in document j, $|T|$ the total number of terms, and $TFIDF(t_i, d_j) = (\text{Term Frequency of Term } t_i \text{ in } d_j) \times \log(\frac{\text{Number of Documents}}{\text{Document Frequency of Term } t_i})$.

A comparison of the three proposed methods with Dona's method [4] is shown in Table 1.

Table 1. A comparison of the three proposed methods with Dona's method

Methods	Dona's method [4]	The first proposed method	The second proposed method	The third proposed method
Feature selection	Mutual Information Measure [2, pp. 139-141]	Mutual Information Measure [2, pp. 139-141]	Mutual Information Measure [2, pp. 139-141]	Formula (4)
Term weight	N/A	Formula (6)	Boolean	Boolean
Similarity measure	N/A	Formula (5)	Formula (7)	Formula (7)
Classifier	Naïve Bayes [24]	k-NN [24]	k-NN [24]	k-NN [24]

In the following, we summarize the three proposed methods as follows:

(A) The First Proposed Method for Text Categorization:

Step 1: Select a predefined number of features based on the Mutual Information (MI) Measure [2] shown in formula (3) to reduce the number of features of each document.
Step 2: Given a testing document, calculate the term weight of the testing document by using formula (6). Find its k-nearest documents among the training documents by using formula (5).
Step 3: The testing document belonging to the category has the largest summing weight.

(B) The Second Proposed Method for Text Categorization:

Step 1: Select a predefined number of features based on the Mutual Information (MI) Measure [2] shown in formula (3) to reduce the number of features of each document.
Step 2: Given a testing document, find its k-nearest documents among the training documents by using the proposed document-document similarity measure described as follows. We use the Boolean method for document representation. Each term weight is either 0 or 1, where 0 means that the term is not appearing and 1 means that it is appearing in the document. Let $M(d_1, d_2)$ denote the number of terms appearing in documents D_1 and D_2, simultaneously. The proposed similarity measure to calculate the degree of similarity $Similarity(d_1, d_2)$ between documents is shown as follows:

$$Similarity(d_1, d_2) = \frac{M(d_1, d_2)}{\sqrt{|d_1| \times |d_2|}}, \tag{7}$$

where $|d_1|$ denotes the number of terms in document d_1 and $|d_2|$ denotes the number of terms in document d_2. Calculate the likelihood of the testing document belonging to each category by summing the weights of its k-nearest documents belonging to the category. For example, assume that there are 3-nearest training documents d_1, d_2, and d_3 of testing document d_4 as shown in Fig. 1. Assume that the degree of similarity between document d_1 and the testing documents d_4 is w_1, the degree of the similarity between document d_2 and the testing document d_4 is w_2, and the degree of similarity between document d_3 and the testing document d_4 is w_3. Then, the summing weights of the documents d_1 and d_2 belonging to Category 1 are equal to $w_1 + w_2$, the weight of d_3 belonging to Category 2 is w_3, if $(w_1 + w_2) < w_3$, then we let the testing document d_4 belong to Category 2.

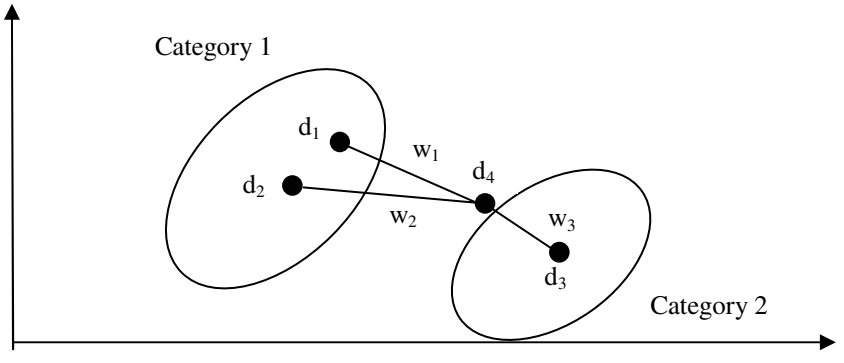

Fig. 1. The 3-nearest training documents of testing document d_4

Step 3: The testing document belonging to the category has the largest summing weight.

(C) The Third Proposed Method for Text Categorization:

Step 1: Select a predefined number of features based on the proposed feature selection method shown in formula (4) to reduce the number of features of each document.
Step 2: The same as **Step 2** of the second proposed method.
Step 3: The testing document belonging to the category has the largest summing weight.

5 Experimental Results

In our experiment, we use the Reuters-21578 "top 10 categories" data set [4], [25] shown in Table 2 for dealing with the text categorization.

Table 2. Top 10 categories of the Reuters-21578 data set [4], [25]

Category	Number of Training Documents	Number of Testing Documents
Earn	2877	1083
Acq	1650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	368	117
Interest	347	131
Ship	197	89
Wheat	212	71
Corn	181	56
Total	7769	3019

We have implemented the proposed method by using MATLAB version 6.5 on a Pentium 4 PC. We use the Microaveraged F_1 [2] for evaluating the performance of the proposed methods. Precision and Recall are defined as follows [2]:

$$\text{Precision} = \frac{\text{number of documents retrieved that are relevant}}{\text{total numberof documents that are retrieved}}, \tag{8}$$

$$\text{Recall} = \frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are relevant}}. \tag{9}$$

The relationship between the precision and the recall is characterized by a graph called the precision-recall curve. The F_1 measure combines the precision and the recall defined as follows [2]:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{10}$$

For multiple categories, the precision, the recall, the microaveraged precision, the microaveraged recall, and the microaveraged F_1 are calculated based on the global contingency matrix shown in Table 3, where

$$p_i = \frac{a_i}{a_i + c_i}, \tag{11}$$

$$r_i = \frac{a_i}{a_i + b_i}, \tag{12}$$

$$\text{microaveraged precision} = \frac{A}{A + C} = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} (a_i + c_i)}, \tag{13}$$

$$\text{microaveraged recall} = \frac{A}{A + B} = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} (a_i + b_i)}, \tag{14}$$

$$\text{microaveraged } F_1 = \frac{2 \times \text{microaverage\_precision} \times \text{microaverage\_recall}}{\text{micoraverage\_precision} + \text{microaverage\_recall}}. \tag{15}$$

Table 3. The Contingency Matrix for a Set of Categories [2]

Category C = {c_1, c_2, … $c_{|c|}$}	Predicted "YES"	Predicted "NO"
Actual class "YES"	$A = \sum_{i}^{\lvert c\rvert} a_i$	$B = \sum_{i}^{\lvert c\rvert} b_i$
Actual class "NO"	$C = \sum_{i}^{\lvert c\rvert} c_i$	$D = \sum_{i}^{\lvert c\rvert} d_i$

In our experiment, we use the proposed three methods (i.e. the first proposed method, the second proposed method and the third proposed method) shown in Table 1 to compare their performance for text categorization with the performance of the method presented in [4]. To compare the proposed three methods with the Dona's Method [4], the microaveraged F_1 is considered for comparing the system performances of different methods for text categorization. Table 4 shows the results of the Retuers-21578 top 10 categories text categorization of the methods. The experimental results show that the proposed three methods (i.e. the first proposed method, the second proposed method and the third proposed method) get higher performances than that of Dona's method [4].

Table 4. A comparison of the performance of categorizing Reuters-21578 top 10 categories data set for different methods

Methods / F_1 measure / Categories	Dona's method [4]	The first proposed method	The second proposed method	The third proposed method
Earn	98.04	96.82	97.76	97.54
Acq	96.67	89.2	95.43	95.56
Money-fx	76.54	73.13	73.5	73.12
Grain	57.47	57.41	60	59.67
Crude	79.43	70.7	73.68	74.42
Trade	85.60	65.63	70.72	72.44
Interest	73.38	60.97	60.76	65.25
Ship	68.75	84.85	85.71	86.75
Wheat	48.39	39.27	44.25	47.01
Corn	44.02	42.95	38.71	37.57
Microaveraged F_1	74.06	81.99	84.6	84.73

6 Conclusions

In this paper, we have presented a new feature selection method based on document frequencies and statistical values. We also have presented a new similarity measure to

calculate the degree of similarity between documents. Based on the proposed feature selection method and the proposed similarity measures between documents, we also have presented three methods to deal with the categorization of the Reuters-21578 top 10 categories data set. The experimental results show that the proposed three methods get higher performance for text categorization than the method presented in [4].

Acknowledgements

The authors would like to thank Professor Yuh-Jye Lee for his help during this research. This work was supported in part by the National Science Council, Republic of China, under grant NSC 94-2213-E-011-003.

References

1. Caropreso, M. F., Matwin, S., Sebastiani, F.: A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization. In: A. G. Chin, eds. Text Databases and Document Management: Theory and Practice, Idea Group Publishing, Hershey, PA (2001) 78–102
2. Chakrabarti, S.: Mining the Web. New York: Morgan Kaufmann (2003) 137–144
3. Chua, S., K, N.: Semantic Feature Selection Using WordNet. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (2004)
4. Doan, S.: An Efficient Feature Selection Using Multi-Criteria in Text Categorization. Proceedings of the IEEE Fourth International Conference on Hybrid Intelligent Systems (2004)
5. Dumais, S. T., Plant, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. Proceedings of the 7th ACM International Conference on Information and Knowledge Management (1998) 148–155
6. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (2000) 59–68
7. Lam, W., Ho, C. Y.: Using a Generalized Instance Set for Automatic Text Categorization. Proceedings of SIGIR-98 the 21st ACM International Conference on Research and Development in Information Retrieval (1998) 195–202
8. Larkey, L. S., Croft, W. B.: Combining Classifiers in Text Categorization. Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (1996) 289–297
9. Larkey, L. S.: Automatic Essay Grading Using Text Categorization Techniques. Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (1998) 90–95
10. Lewis, D. D.: An evaluation of phrasal and clustered representations on a text categorization task. Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (1992) 37–50
11. Lewis, D. D.: Representation and Learning in Information Retrieval. Ph.D. Dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA (1992)
12. Lewis, D. D.: and Ringuette, M., A Comparison of Two Learning Algorithms for Text Categorization. Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (1994) 81–93

13. Li, Y. H., Jain, A. K.: Classification of Text Documents, Computer Journal Vol. 41, No. 8 (1998) 537–546
14. Li, H., Yamanishi, K.: Text Classification Using ESC-Based Stochastic Decision Lists. Proceedings of the 8th ACM International Conference on Information and Knowledge Management (1999) 122–130
15. Mladenic, D.: Feature Subset Selection in Text Learning. Proceedings of the 10th European Conference on Machine Learning (1998) 95–100
16. Ng, H. T., Goh, W. B., Low, K. L.: Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization. Proceedings of the 20th ACM International Conference on Research and Development in Information Retrieval (1997) 67–73
17. Porter, M. F.: An Algorithm for Suffic Stripping Program. Vol. 14, No. 3 (1980) 130–137
18. Salton, G., Wong, A., Yang, C.: A Vector Space Model for Automatic Indexing. Communications of the ACM, Vol. 18, No. 11 (1975) 613–620
19. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Survey, Vol. 34, No. 1 (2002) 1–47
20. Sebastiani, F., Sperduti, A., Valdambrini, N.: An Improved Boosting Algorithm and its Application to Automated Text Categorization. Proceedings of the 9th ACM International Conference on Information and Knowledge Management (2000) 78–85
21. Shima, K., Todoriki, M., Suzuki, A.: SVM-Based Feature Selection of Latent Semantic Features. Pattern Recognition Letters 25 (2004) 1051–1057
22. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval Journal, Vol. 1, No. 1–2, (1999) 69–90
23. Yang, Y. Pedersen, J.: A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning (1997) 412–420
24. Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. Proceedings of the SIGIR-99 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA (1999) 42–49
25. Reuter-21578 Apte Split Data Set, http://kdd.ics.uci.edu/data-bases/reuter21578/reuter221578.html