An Improved Algorithm of Bayesian Text Categorization

Tao Dong and Wenqian Shang
School of Computer, Communication University of China, Beijing, China
Email: {victor666.ok, shangwenqian}@163.com

Haibin Zhu
Dept. of Computer Science, Nipissing University, North Bay, Canada
Email: haibinz@nipissingu.ca

Abstract—**Text categorization is a fundamental methodology of text mining and a hot topic of the research of data mining and web mining in recent years. It plays an important role in building traditional information retrieval, web indexing architecture, Web information retrieval, and so on. This paper presents an improved algorithm of text categorization that combines the feature weighting technique with Naïve Bayesian classifier. Experimental results show that using the improved Gini index algorithm to feature weight can improve the performance of Naïve Bayesian classifier effectively. This algorithm obtains good application in the sensitive information recognition system.**

Index Terms—**text categorization, Gini index, feature weighting, Naïve Bayes**

I. INTRODUCTION

A. Background

With the rapid development of network technologies, the network data grow exponentially. How to effectively organize and manage information and quickly, accurately and completely find useful information for users is a major challenge for information sciences and technologies. As a key technique to process and organize large amount of texts, text categorization can solve the problem of information clutters to a large extent, and make users locate the information they need rapidly and accurately. Therefore, text categorization has become a fundamental technology with great practical values and is well-accepted and has made great progresses [1][2].

Feature selection is an important step of text categorization. The strategy of feature selection is to select a specific amount of useful features for categorization, and delete the rest useless features completely. The feature selection in classical Bayesian classifier is helpful to improve the accuracy of categorization to some extent, but it treats the remained features evenly. Obviously, different features have different influences on the result of categorization, hence we need to give different weights to different features.

There are some commonly used algorithms of text categorization: kNN, Naïve Bayes, SVM, neural networks, maximum entropy and so on. Among them, naive Bayesian classifiers get the extensive attentions and universal applications with their unique advantages of high speed, small error rate and implementations online.

Therefore, this paper presents a Naive Bayesian classifier based on an improved feature weight algorithm of Gini index. Experimental results show that our method is effective and feasible.

B. The Research Status of Text Categorization

Abroad, the research on text categorization began in the late 1950s, Luhn pioneered this field by using the thought of word frequency statistics into text categorization. In 1960, Maron published the first papers about automatic categorization algorithm. Then, K. Spark, G. Salton, KS Jones and many other scholars also made very effective work in this field of research. Now the research on text categorization abroad have been entered from the experimental stage to the practical stage, and achieved a wide range of applications in the mail categorization, electronic conference and so on. Among them, the e-mail categorization system for the White House which developed by Massachusetts Institute of Technology and the construe system for Reuters which developed by the Carnegie Group are more successful [3].

Compared with the English text, Chinese text categorization has an important difference in the pre-processing stage. Unlike the English words which distinguish by spaces, Chinese texts need segmentation. Thus, the Chinese text categorization mainly focuses on how to use some features of Chinese themselves to represent the whole text better. Although the domestic research for text categorization starts late, Chinese segmentation technology has become mature from a simple dictionary approach to the segmentation based on statistical language model.

In 1981, Professor Hanqing Hou discussed and elaborated computer's application in the text categorization. Since then, our country produces a number of text categorization systems, including representatives of a Chinese automatic categorization system based on neural network algorithm developed by Shanghai Jiao Tong University and an automatic categorization system of Tsinghua University. At the

doi:10.4304/jsw.6.9.1837-1843

same time, the domestic scholars also carry out extensive research and implementation in different categorization algorithms. Xiaoli Li and Zhongzhi Shi of CAS Institute of Computing apply the concept inference network for text classification and get the recall of 94.2% and accuracy of 99.4% [4]. Zhong Fan of University of Science and Technology of China proposes a Hypertext Coordination Classifier based on KNN, Bayesian and document similarity and gets the accuracy of nearly 80% [5]. It is appropriate to consider the structured information of HTML text. Xuanjing Huang and Lide Wu of Fujitsu Research Center and Fu Dan University study the text categorization of independent language, using the mutual information of vocabulary and class for the score function, considering the single-categorization and multi-categorization and get the best recall of 88.87% [6]. Qian Diao and Yongcheng Wang of Shanghai Jiao Tong University combine the term weight with algorithm to make categorization and get the accuracy of 97% in a closed testing experiment based on VSM [3].

C. New Development of Text Categorization

In recent years, text categorization has become a popular topic for a number of researchers in many areas. The researchers introduce more and more knowledge to the field of text categorization from different perspectives, promote the continuous development of text categorization and invent many new ways such as text categorization model based on the fuzzy-rough, fusion of multiple classifiers, latent semantic categorization model, text categorization model based on the RBF network and so on.

II. Classical Naïve Bayesian Classifier

Naive Bayesian classifiers assume that the value of each feature has an independent influence on a given class, and this assumption called class conditional independence that used to simplify the computation, and in this sense, we call it "Naive".

Bayesian method is a commonly used supervised categorization algorithm. It is a kind of pattern recognition method based on Bayes theorem that known prior probability and conditional probability. Therefore, we first introduce the probability basis of Bayesian classifier.

A. The basis of Bayesian probability

1) Prior probability

Prior probability is based on historical data or subjective judgments to determine the probability of each event. Because this kind of probability is a pre-test probability and can't be confirmed through experiments, we called it priori probability. Prior probability is generally divided into two types of objective and subjective prior probability. Objective prior probability refers to use historical data to calculate the probability, and subjective prior probability refers to use people's experience to determine the probability when the historical data is absent or incomplete.

2) Posteriori probability

Posteriori probability generally refers to use Bayes formula and other means like survey to obtain new additional information. It is a more realistic probability by amending the prior probability.

3) Joint probability

Joint probability is also called multiplication formula, it is the probability of the product of two arbitrary events or the probability of cross-event.

4) Total probability formula

If all the factors$(B_1', B_2', \ldots)$ that affect A' meet $B_i' \cdot B_j' = \varphi$, $(i \neq j)$, and $\sum_{i=1}^{t} P(B_i') = 1$, It certainly has:

$$P(A') = \sum P(B_i')P(A'|B_i') \tag{1}$$

5) Bayesian formula

Bayesian formula is also called the posteriori probability formula or inverse probability formula.

If the prior probability is $P(B_i')$, and the new additional information obtained by investigation is $P(A_j'|B_i')$, among it, i=1,2,…z, j=1,2,… z′. Then the posteriori probability calculated by the Bayesian formula is:

$$P(B_i'|A_j') = \frac{P(B_i')P(A_j'|B_i')}{\sum_{t=1}^{z'} P(B_i')P(A_t'|B_i')} \tag{2}$$

B. Bayesian Theorem

We assume that d is a data sample with unknown class label and H′ is an assumption. If data sample d belongs to a particular class c, for the problem of categorization, we hope to get $P(H'|d)$. Namely, we hope to know the probability of H′ when data sample d is given.

$P(H'|d)$ is a posteriori probability or a posteriori probability under the condition of d. $P(H')$ is a prior probability or a prior probability of H′, and it is independent of d.

Similarly, $P(d|H')$ is a posteriori probability of d under the condition of H′ and P(d) is a prior probability of d.

But how can we calculate these probabilities? As described below, P(d), $P(H')$ and $P(d|H')$ can be calculated from the given data. The Bayesian theorem provides a method for calculating the posteriori probability by P(d), $P(H')$ and $P(d|H')$. So, Bayesian theorem can be described as follows [7]:

$$P(H'|d) = \frac{P(d|H')P(H')}{P(d)} \tag{3}$$

Each data sample is represented as an n-dimensional feature vector that describes n measures of n samples.

Assumed m classes of c_1, c_2,…c_m and given an unknown data sample d (no class label), they will be sorted into the class which has the highest posteriori probability based on categorization. In other words, a naive Bayesian classifier will assign unknown samples to the class c_i, if and only if: $P(c_i|d) > P(c_j|d)$, $1 \le i, j \le m$, $j \neq i$.

Thus, we can maximize the $P(c_i|d)$, where class c_i has the largest $P(c_i|d)$ and is called the maximum posteriori assumption. According to Bayesian theorem (1):

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{4}$$

Since P(d) is a constant for all classes, we only need to maximize $P(d|c_i)P(c_i)$[8]. If the prior probability of the class is unknown, it is usually assumed that the probability of these classes is equivalent, that is $P(c_1) = P(c_2) = \cdots = P(c_m)$. So we maximize $P(d|c_i)$ only. Otherwise, we should maximize the $P(d|c_i)P(c_i)$. Please note that, the prior probability of a class can be calculated by $P(c_i) = \frac{s_i}{s}$, where s_i is the number of training samples of the class and s is the total number of training samples.

It may cost too much to calculate $P(d|c_i)$ when the given data sets with many attributes. To reduce the computational cost of $P(d|c_i)$, we can simply assume that the class is conditional independent. If we know the class label of a sample, and assume that the value of each property is conditional independent, namely, there is no dependent relationship between every pair of properties. Hence:

$$P(d|c_i) = \prod_{j=1}^{n} P(x_j|c_i) \quad (5)$$

C. Process of Naïve Bayesian Categorization

$P(x_1|c_i)$, $P(x_2|c_i)$,…, $P(x_n|c_i)$ can be valued by training samples. Moreover:

If E_j is a classified property, we get $P(x_j|c_i) = \frac{s_{ij}}{s_i}$, where s_{ij} is the number of training samples of class c_i with the value of x_i based on the property E_j and s_i is the number of training samples of c_i.

If E_j is continuous-valued attribute, we usually assume that the properties obey the Gaussian distribution. Thus,

$$P(x_j|c_i) = g(x_i, u_{c_i}, \sigma_{c_i}) = \frac{1}{\sqrt{2\pi}\sigma_{c_i}} e^{\frac{(x_j - u_{c_i})}{2\sigma_{c_i}^2}} \quad (6)$$

Where, $g(x_i, u_{c_i}, \sigma_{c_i})$ is the Gaussian density function of the property E_j', u_{c_i} is the mean and σ_{c_i} is the standard deviation.

In order to classify the unknown sample d and calculate $P(d|c_i)P(c_i)$ of each class c_i, we assign sample d to class c_i, if and only if:

$$P(d|c_i)P(c_i) > P(d|c_j)P(c_j),\ 1 \le i, j \le m,\ j \ne i \quad (7)$$

In other words, d is assigned to class c_i with the largest $P(d|c_i)P(c_i)$.

As for the estimation of the probability, the m-estimate or Laplace estimate can improve the reliability of estimates. Hence, we use the Laplace estimate and the formula is as follows:

$$P(c_j) = \frac{\sum_{i=1}^{|D|} P(c_j|d_i)}{|D|},\ j=1,2,\ldots|C| \quad (8)$$

$$P(d_i|c_j) = \frac{1+\sum_{i=1}^{|D|} B_{it} P(c_j|d_i)}{2+\sum_{i=1}^{|D|} P(c_j|d_i)},\ j=1,2\ldots|C|;t=1,2\ldots \quad (9)$$

Where D is the training text set and $P(c_j|d_i) \in \{0,1\}$ means whether the training document d_i belongs to the class c_j, where 1 means yes but 0 no.

There are mainly two kinds of naive Bayesian models for different implementations. One is the multivariate Bernoulli model that only considers whether the feature item appears in the text, if the feature item appears, denoted as 1, otherwise as 0. The other is the multinomial model that considers the characteristics' number of occurrences in the text.

In the multivariate Bernoulli model,

$$P(d|c_j) = \prod_{t=1}^{n}(B_{xt} P(x_t|c_j) + (1 - B_{xt})(1 - P(x_t|c_j))) \quad (10)$$

Where x_t is the t^{th} characteristic (the t^{th} component of the vector of text), n is the total number of features and B_{xt} represents whether the feature x_t appears in text d. Besides,

$$P(x_t|c_j) = \frac{1+c_{jt}}{2+N_{c_j}} \quad (11)$$

Where C_{jt} is the number of texts in C_j that contains feature x_t, and N_{cj} the number of all texts in C_j. In the multinomial model,

$$P(d|c_j) = \prod_{t=1}^{n} \frac{P(x_t|c_j)^{N_{it}}}{N_{xt}!} \quad (12)$$

Where N_{xt} represents the number of occurrences in the text of feature x_t,

$$P(x_t|c_j) = \frac{1+\sum_{t=1}^{n} N_{it} P(c_j|d_i)}{|V|+\sum_{t=1}^{|V|}\sum_{t=1}^{n} N_{it} P(c_j|d_i)} \quad (13)$$

Where |V| represents the number of occurrences in the text d_i of feature x_t.

For the no-label text in the test texts, we can use the trained classifiers to find the posteriori probability of text d which belongs to class c_j. We use x_t to represent t^{th} characteristic entry in text d, and the formula as follows:

$$P(c_j|d) \propto P(c_j) \prod_{t=1}^{n} P(x_t|c_j) \quad (14)$$

In this paper, we choose the multinomial model.

III. IMPROVED NAÏVE BAYESIAN CLASSIFIER

A. Traditional Gini Index Algorithm

Gini index is a non-purity method to split properties for classes, binary, discrete and other types of fields. It is proposed by Breiman et al. in 1984 [9] and has been widely used in the CART algorithm, SLIQ algorithm, SPRINT algorithm and the decision tree algorithm of the Intelligent Miner algorithm. The algorithm is described as follows:

We assume that Q is a set of data samples of s, its class labels have m different values which define m different classes (c_i, i=1…m). |C| is the total number of classes, and we divide Q into m sub-sets according to class labels (Q_i, i=1…m). We assume that Q_i is a set of samples belonging to class c_i, s_i is the number of samples in Q_i. Then the Gini index of Q is:

$$Gini(Q) = 1 - \sum_{i=1}^{|C|} P_i^2 \quad (15)$$

Where P_i is the probability of any sample belonging to c_i that estimated by s_i/s. When Gini(Q) is the minimum 0, namely, all the samples in the set belong to the same class and we can get the maximum useful information at this time; when all samples in the set have a uniform distribution for the classes, the Gini (Q) get the maximum

value and we get the minimum useful information at this time.

B. Improved Gini Index Algorithm

S. Shankar and G. Karypis [10] studied the application of the Gini index in feature weighting of the categorization by centroid. They used a time-consuming iterative method, which focused on feature weighting, and did not discuss the feature selection. Charu C. Aggarwal [11] studied the Gini index on feature selection of text categorization, but they used the Gini index of the hybrid degree. Our method is completely different from their methods, we construct a new measure function of Gini index through in-depth analysis of the Gini index and texts' features and complete the feature selection in the original feature space. We use the Gini index of purity for not only the categorization by centroid but also other categorization methods.

The initial form of Gini index is to measure a "hybrid degree", i.e., the property for categorization, namely, the smaller "hybrid degree" the better property. If we use the following form [12][13]:

$$\mathrm{Gini}(W) = \sum_{i=1}^{|C|} P_i^2 \tag{16}$$

It is to measure a "purity" that is the property for categorization, namely, the larger "purity" the better property. In literature [14], they also use the "purity" measured form of Gini. This form helps to reflect the impact of feature selection on categorization, hence we also use this measurement to conduct the feature selection of texts.

This "purity" form of the Gini index can be further changed as follows:

$$\mathrm{Gini}(W) = \sum_{i=1}^{|C|} \sqrt{P(W|C_i)} \tag{17}$$

C. Feature Weighting Technique

Feature weighting has the following three general steps:

(1) Calculating the ability of distinguish for each feature; (2) Screening a certain number of features according to the ability to distinguish; (3) Adjusting the weights of features, emphasizing the features with a strong ability to distinguish, and inhibiting the lower or no one.

Step (1) is to calculate the ability of identification for each feature by constructing a feature evaluation function (ie feature selection function). The commonly used evaluation function is extended from information theory, such as Information Gain, Expected Cross Entropy, Mutual Information, Odds Ratio, Term Strength, etc. It is used to mark each feature and has a good reflection of the feature and the degree of the correlation between features and classes. The ability of identification for each feature is measured by the assessment point.

There are two ways to execute Step (2): Method 1, setting a threshold of assessment and deleting the features below the threshold; Method 2, setting a threshold of retained number of features, sorting the features by the assessment and retaining the top predetermined number of features.

Both methods have their advantages and disadvantages. Method 1 has the advantage of no sorting algorithm and high time efficiency, but it is difficult to determine the threshold as it is related to the evaluation function, besides, it is also changeable with the change of the training samples. Method 2 is better to determine the threshold, but it must sort the assessment point. In this way, the time complexity is also O (nlog n) even with fast sorting method, where n is the total number of features of the training samples.

Step (3) is to construct a strategy to adjust the weight. Weight adjustment aims to highlight important features and inhibit the secondary ones [15].

D. TF-IDF Algorithm

TF-IDF algorithm was first proposed by Salton and Buckley in 1988 and used for information retrieval. Then it was applied for feature weighting in data mining such as text categorization and clustering. It calculates the feature's weight in the text based on its Term Frequency and Inverse Document Frequency.

We suppose N is the total number of texts in the training samples, df_i is the number of the text which containing feature t_i and f_{ij} is the number of feature t_i which appearing in text d_j. So the Term Frequency (defined as tf_{ij}) of t_i in text d_j is given as follows:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, f_{3j}, \cdots, f_{|V|j}\}} \tag{18}$$

In the above formula, the denominator is the maximum value of f_{ij}. If t_i doesn't appear in the text d_j, then $f_{ij} = 0$. $|V|$ is the total number of features in the training samples [16][17].

The Inverse Document Frequency (defined as idf_i) of t_i is given as follows:

$$idf_i = \log \frac{N}{df_i} \tag{19}$$

So the final TF-IDF weight is given below:

$$w_{ij} = tf_{ij} \times idf_i \tag{20}$$

It can be seen from this formula that the more time a feature appears in a text the higher weight it will get. And a feature appears in the more texts, it will get the less importance. This method is effective for information retrieval but not for text categorization and clustering. For text categorization and clustering, a feature with higher document frequency is more important than the lower one, which is opposite for the review in information retrieval. In addition, TF-IDF only represents the feature's ability of distinguishing a text but not contain its ability of distinguishing a class and other classes. But for text categorization and clustering, a feature's distinct for class is more important. So the original IDF is inappropriate for text categorization and clustering.

Therefore, we use a feature evaluation function to replace the IDF function and construct a new feature weight function, TF-TWF function. TWF represents a feature evaluation function, the TF-TWF weighting formula is as follows:

$$W_t = TF - TWF(x_t) = TF(x_t) \times TWF(x_t) \quad (21)$$

Among them, $TF(x_t)$ means the word frequency of feature t in text d. $TWF(x_t)$ is a common evaluation function that is used to mark each feature and reflects the correlation between features and various types.

After the weight adjustment based on TF-TWF, the feature's importance in the classifier has changed with the change of weight. According to the adjusted feature's weight, modifying the feature's importance in the classifier, then we can calculate the $P(c_j|d)$ as follows:

$$P(c_j|d) = \log[P(c_j)] + \sum_{t=1}^{n} TF - TWF(x_t) \times \log[P(x_t|c_j)] \quad (22)$$

Where $TF\text{-}TWF(x_t)$ is a new weight function of feature x_t. The feature that has a higher weight plays a greater role in the naive Bayesian classifier; and the feature with a smaller $TF\text{-}TWF(x_t)$ plays a smaller role in the naive Bayesian classifier[18].

E. The New Bayesian Decision Model

So we design a new feature weighting function, namely, TF-Gini function. We use the Gini Index to replace the IDF in our improved algorithm. Through the description above, we can get the new Bayesian decision model as follows:

$$P(c_j|d) = \log[P(c_j)] + \sum_{t=1}^{n} TF - Gini(x_t) \times \log[P(x_t|c_j)] \quad (23)$$

Then the new decision rule of our improved Naive Bayesian classifier is assigning d to the class of the maximum probability $P(c_j|d)$, namely, getting the $\arg\max P(c_j|d)$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The data of experiment1 and experiment2 comes from the articles in a large number of Chinese websites. These data include two classes of sensitive information and non-sensitive information.

Experiment1 uses 1500 texts. The set of training samples has 1000 texts which consist of 500 texts about sensitive information and 500 texts about non-sensitive information; the set of test samples has 500 texts which consist of 250 texts about sensitive information and 250 texts about non-sensitive. There is no overlap between the training samples and the test samples. The feature selection for reservations is 2000.

The experimental results are as follows:

Table 1. Comparison of categorization performance

Algorithm	Precision (%)	Recall (%)	F-score (%)
kNN	96.76	55.41	70.44
Naïve Bayes	96.60	96.60	96.60
Improved Naïve Bayes	83.61	100.00	91.08

Experiment2 uses 1000 texts. The set of training samples has 900 texts which consist of 450 texts about sensitive information and 450 texts about non-sensitive information; the set of test samples has 100 texts which consist of 50 texts about information and 50 texts about non-sensitive information. There is no overlap between the training samples and the test samples. The feature selection for reservation is 2000.

The experimental results are as follows:

Table 2. Comparison of categorization performance

Algorithm	Precision (%)	Recall (%)	F-score (%)
kNN	69.20	72.00	70.60
Naïve Bayes	80.77	84.00	82.35
Improved Naïve Bayes	71.88	92.00	80.70

The data of experiment3 comes from Fu Dan standard Chinese corpus. We choose three classes from this corpus which include 386 texts. The set of training samples has 190 texts which consist of 59 texts about Education, 74 texts about Military and 57 texts about Transport; the set of test samples has 196 texts which consist of 61 texts about Education, 76 texts about Military and 59 texts about Transport. There is no overlap between the training samples and the test samples. The feature selection for reservation is 2000.

The experimental results are as follows:

Table 3. Comparison of categorization performance

Algorithm	Precision (%)	Recall (%)	F-score (%)
Naïve Bayes	88.89	94.92	91.80
Improved Naïve Bayes	77.33	98.30	86.57

From Figure 1 and Figure 2, we can see that the improved Naive Bayesian classifier has shown better results on the different sensitive information data sets. It increases 10 to 20 percent on the categorization performance compared to the kNN classifier. This improvement is obvious. Although slightly inferior to the Naive Bayesian classifier on the accuracy of categorization, the recall has been increased 4 to 8 percent. As we know, in an identification system for sensitive information, the most important performance indicator is identifying the sensitive information as much as possible and not missing sensitive information. In other words, the value of this system is mostly determined by the recall. In this regard, the improved algorithm has achieved a great success.

From Figure 3, we can see that the improved Naive Bayesian classifier has also shown better results on the multi-categorization. As kNN classifier is totally inappropriate for multi-categorization, we just compare the improved Naïve Bayesian classifier with the Naïve Bayesian classifier. The results show that the improved algorithm increases nearly 4 percent on the recall relative to the original one. The recall means the accuracy of the information that users interested to. In this regard, the improved algorithm has a greater practical value for users.

V. CONCLUSION

This paper introduces a new algorithm of text categorization--- Naïve Bayesian classifier based on the improved feature weighting algorithm of Gini index. The algorithm takes into account that different features have different usefulness for categorization, integrates the Gini index to feature weighting techniques effectively and gives the feature different weights. At last, we combine it with the Naive Bayesian classifier. Experimental results show that this algorithm has a good performance on the accuracy and practical value in the sensitive information recognition system, therefore this improvement is successful.

ACKNOWLEDGMENT

This research is partly supported by the project "Digital New Media Content Production, Integration, Operation and monitoring (2009)" of Beijing Municipal Special Fund for Cultural and Creative Industries and partly supported by the project "Engineering planning project of Communication University of China"(XNG1030) and partly supported by 2010 National Science and Technology Support Program(2009BAH40B04) and partly supported by 48th group of post doctor fund of china(20100480357).

REFERENCES

[1] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys,2002, 34(1), pp. 1-47.
[2] J. Carbonell, W. W. Cohen, Y. Yang. Guest Editorial for the Special Issue on Text Categorization. Machine Learning, 2000.
[3] S.Q. Gao. The Review of Research Status on Web Text Categorization. Book, Information and Knowledge, 2008, No.123, pp. 81-86.
[4] X.L. Li. Concept Reasoning Network and its Application in Text Categorization. Computer Research and Development, 2000(9).
[5] B. Zhang. The Research of Chinese Text Categorization. Master Thesis of Wuhan University, 2004.
[6] X.J. Huang, L.D. Wu. The Text Categorization Based On the Independent of the Language. International Conference on Multilingual Information Processing, 2000, pp. 37-43.
[7] Chia-Hung Yeh, Kai-Jie Fan, Mei-Juan Chen, Gwo-Long Li. Fast Mode Decision Algorithm for Scalable Video Coding Using Bayesian Theorem Detection and Markov Process. IEEE Transaction on Circuits and Systems for Video Technology, April 2010, vol.20, 4, pp.563-574.
[8] F. Li, Q.S. Liu.Naive Bayesian classifier model based on improved weighted attributes. Computer Engineering and Applications, 2010, 46(4), pp.132-133.
[9] L. Breiman, J. Friedman, R. Olshen et al. Classification and Regression Trees. Monterey, CA: Wadsworth International Group, 1984.
[10] S. Shankar, G. Karypis. A Feature Weight Adjustment Algorithm for Document Categorization. http://www.cs.umm.edu/~karypis.
[11] Charu C. Aggarwal, Stephen C. Gates, Philip S. Yu. On the Merits of Building Categorization Systems by Supervised Clustering. In KDD'99, San Diego, USA, 1999, pp. 352-356.
[12] W.Q. Shang, H.K. Huang, Y.L. Liu, Y.M. Lin, Y.L. Qu, H.B. Dong. Research On the Algorithm of Feature Selection Based on Gini Index for Text Categorization. Computer Research and Development, 2006, 43(10), pp.1688-1694.
[13] H. Park, S. Kwon, H. Kwon. Complete Gini-Index Text (GIT) feature selection algorithm for text classification. 2010 2nd International Conference on Software Engineering and Data Mining (SEDM), 2010, pp.366-371.
[14] S. K. Gupta, D. V. Somayajulu, J. K. Arora, B. Vasudha. Scalable Classifiers with Dynamic Pruning. In Proc. of the 9th International Workshop on Database and Expert Systems Applications. Los Alamitos, CA: IEEE Computer Society Washington, DC, USA, 1998.
[15] H.L. Tang, J.T. Sun, Y.C. Lu. Text categorization and evaluation function in combination with TEF-WA weight adjustment technique. Computer Research and Development, 2005, 42(1), pp.47-53.
[16] P. Castells, M. Fernandez, D. Vallet. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering, February 2007, vol.19, 2, pp.261-272.
[17] H.K. Mohamed. Automatic documents classification. International Conference on Computer Engineering & Systems, 2007, pp.33-37.
[18] D.Y. Wang, J. Wang. Improved Feature Weighting Algorithm for Text Categorization. Computer Engineering, 2010, 36(9), pp.197-199.

Tao Dong Born in Mianyang of Sichuan Province, China, in 1988. He received the bachelor's degree of electronic and information engineering in Zhejiang Gongshang University, Hangzhou, China in June 2010. Currently, he is a Master candidate of computer technology in Communication University of China, Beijing, China. His main research interests Text Mining, Web Mining and Information Retrieval.

Wenqian Shang Born in Kaifeng of Henan Province in 1971. She received her bachelor's degree in school of computer, Southeast University, Nanjing China in June 1994; received her MS degree in school of computer, National University of Defense Technology, Changsha China in June 1999; received her Ph.D. degree in school of computer, Beijing Jiaotong University, Beijing China in January 2008. Her major field of study is AI and text mining.

She is now an associate professor of computer science, Communication University of China, teaching courses in artificial intelligence and research work. Her research interests include Text Mining, Web mining, Natural Language Processing, AI, etc.

Dr. Shang is a member of China computer Federation. She is the chair of ICS workshop, CSO 2009, 2010, 2011. She is the reviewer of ICCASM2011, ICCSIT2011, ICNECS2011, etc.

Haibin Zhu Ph.D., computer science, Changsha Inst. Of Tech.(CIT), Changsha, China, 1997. M.S., computer science, Changsha Inst. Of Tech.(CIT), Changsha, China, 1988.

Now he is Full Professor of the Department of Computer Science and Mathematics, Director and Founder of Collaborative Systems Laboratory, Nipissing University, Canada. His main research interests Role-Based Collaboration, Computer-Supported Cooperative Work, Human Computer Interaction and Service Computing and Cloud Computing.

Dr. Zhu is a senior member of IEEE, a member of ACM, and a life member of the Chinese Association for Science and Technology, USA. He is the recipient of the 2006-2007 research award from Nipissing University, the 2004 and 2005 IBM Eclipse Innovation Grant Awards, the Best Paper Award from the 11th ISPE Int'l Conf. on Concurrent Engineering (ISPE/CE2004), the Educator's Fellowship of OOPSLA'03, a 2nd Class National Award of Excellent Textbook from the Ministry of Education of China (2002), a 2nd class National Award for Education Achievement(1997), and three 1st Class Ministerial Research Achievement Awards from The Commission of Science Technology and Industry for National Defense of China (1997, 1994, and 1991).