Feature Selection Based on Term Frequency and T-Test for Text Categorization

Deqing Wang
SKLSDE, Beihang University
Beijing, 100191, China
dqwang@nlsde.buaa.edu.cn

Hui Zhang
SKLSDE, Beihang University
Beijing, 100191, China
hzhang@nlsde.buaa.edu.cn

Rui Liu, Weifeng Lv
SKLSDE, Beihang University
Beijing, 100191, China
{liurui,lwf}@nlsde.buaa.edu.cn

ABSTRACT

Much work has been done on feature selection. Existing methods are based on document frequency, such as Chi-Square Statistic, Information Gain etc. However, these methods have two shortcomings: one is that they are not reliable for low-frequency terms, and the other is that they only count whether one term occurs in a document and ignore the term frequency. Actually, high-frequency terms within a specific category are often regards as discriminators.

This paper focuses on how to construct the feature selection function based on term frequency, and proposes a new approach based on t-test, which is used to measure the diversity of the distributions of a term between the specific category and the entire corpus. Extensive comparative experiments on two text corpora using three classifiers show that our new approach is comparable to or or slightly better than the state-of-the-art feature selection methods (i.e., χ^2, and IG) in terms of macro-F_1 and micro-F_1.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

feature selection, term frequency, t-test, text classification

1. INTRODUCTION

Text classification (TC) is to assign new unlabeled natural language documents to predefined thematic categories [13]. Many classification algorithms have been proposed for TC, e.g., k-nearest neighbors [20], centroid-based classifier [7], and support vector machines (SVMs) [3].

Generally, text feature space is often sparse and high-dimensional. For instance, the dimensionality of a moderate-sized text corpus can reach up to tens or hundreds of thousands. The high dimensionality of feature space will cause the "curse of dimensionality", increase the training time, and

affect the accuracy of classifiers [13, 6, 20]. Therefore, feature selection techniques are proposed to reduce the dimensionality under the premise of guaranteeing the performance of classifiers. Existing feature selection methods are based on statistical theory and information theory, such as χ^2, IG, MI, and ECE. The theoretical basis of the four methods is sound, but the performances of these methods on TC tasks are different. Both χ^2 and IG often achieved better accuracy than MI and document frequency (DF) [20]. However, other authors suspected the performance of IG on skewed text corpora [11].

Besides the classical methods, many improved methods have been proposed. For example, Yang et al. [19] considered the terms whose relative term frequency was larger than a predefined threshold λ, and then modified the IG formula to select features. Forman [5] proposed the Bi-Normal Separation (BNS) method, which used the standard Normal distribution's inverse cumulative probability function to construct feature selection function. Uguz [15] proposed a two-stage feature selection method for TC by combining IG, principal component analysis and genetic algorithm. More and more methods have been generated, such as, mr2PSO [16], and improved TFIDF method [17]. It is worth noting that t-test has been used for gene expression and genotype data [14, 21]. However, the variable in gene expression or genotype data is different from that in text data, i.e., the term frequency. Thus we try to validate the role of t-test in text feature selection.

From document frequency perspective, the above methods almost use DF sufficiently. However, no efficient method is proposed from term frequency perspective. It inspires our motivation of this paper. Our paper makes the following contributions:

(1) Using central limit theorem (CLT), we prove that the frequency distribution of a term within a specific category or within the entire collection will be approximately normally distributed.

(2) We model the diversity of the frequency of a term between the specific category and the entire corpus with t-test. It means that if the distribution of one term within the specific category is obviously different with that within the entire corpus, the term can be considered to be feature.

(3) We verify our new approach on two common text corpora with three well-established classifiers. The experiments show that our approach is comparable to or even slightly better than the state-of-the-art χ^2 and ECE in terms of both macro-F_1 and micro-F_1, and it outperforms IG and MI methods significantly on unbalanced text corpus.

2. FEATURE SELECTION METRICS

Many feature selection approaches have been proposed in TC tasks, but we only give detailed analysis on four methods because they have been widely used and achieved better performance, the formulae can be found in Refs [20, 5, 6]. They are: Chi-Square Statistic (χ^2), Information Gain (IG), Mutual Information (MI), and Expected Cross-Entropy (ECE).

χ^2 was proposed by Pearson early in 1900 [20]. The χ^2 statistic is used to measure the lack of independence between t_i and C_j, and can be regards as the χ^2 distribution with one degree of freedom. In real-world corpus, χ^2 statistic is based, however, on several assumptions that do not hold for most textual analysis [4]. For instance, if term t_1 occurs in 50% documents of a specific category C_j and term t_2 occurs in 49% documents, but the frequency of t_2 is much higher than that of t_1. Experts often think term t_2 should have more discriminating power than t_1 in the specific category C_j. χ^2, however, will be prone to select term t_1 as feature, rather than t_2. The problem is that χ^2 is not reliable for low-frequency terms [4].

The weakness of MI is that the score is strongly influenced by the marginal probabilities of terms, because rare terms will have a higher score than common terms. Therefore, the scores are not comparable across terms of widely differing frequency [20, 9]. Besides, MI gives longer documents higher weights in the estimation of the feature scores.

IG was firstly used as attribute selection measure in decision tree [20]. This measure is from entropy in information theory, which studies the value or "information content" of messages. IG is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on term t_i). IG is also called average mutual information. The weakness of IG method is that it prefers to select terms distributed in many categories, but these terms have less discriminating power in TC tasks. Differing from IG, Expected Cross-Entropy (ECE) [8] only considers the terms occurred in a document and ignores the absent terms.

As we know, if a term (except stop words) occurs frequently within a specific category, the term should be considered as a feature or discriminator of the category. For example, "computer" occurs frequently in the IT category. However, the above methods are all based on document frequency, and **ignore** the term frequency. In next section, we will propose a new approach based on term frequency, and it can capture the information of high-frequency terms.

3. NEW APPROACH BASED ON TERM FREQUENCY AND T-TEST

The t-test, namely the student t-test, is often used to assess whether the means of two classes are statistically different from each other by calculating a ratio between the difference of two class means and the variability of the two classes [21]. In this section, we explain why the averaged term frequency within a single category or in the whole corpus is approximately normal using Lindeberg-Levy central limit theorems, and then how the t-test is constructed based on the averaged term frequencies.

Let us consider the term frequency in text corpus consisting of n documents. Given a vocabulary V, the term frequency (tf_{ij}) of a term t_i $(1 \leq i \leq |V|)$ in the jth docu-

ment $(1 \leq j \leq N)$ can be considered as a random variable, which subjects to some unknown distribution, e.g., multinomial model [10]. In the multinomial model, a document is an ordered sequence of word events drawn from the same vocabulary V, and the probability of each word event in a document is independent of the word's context and position in the document. Therefore, each document d_j is drawn from a multinomial distribution of words with as many independent trials [10]. That is, the occurrence of one term in each document is dominated by a multinomial function. Then,

(1) Let $\{tf_{i1}, \cdots, tf_{iN}\}$ be a random sample of size N, where N is the number of documents in the collection, and $tf_{ij}(0 \leq j \leq N)$ is the term frequency of t_i in jth document. That is, a sequence of independent and identically distributed random variables with expected values $\mu_i = Np_i$ and variances $\sigma_i^2 = Np_i(1 - p_i)$, where p_i is the distributed probability of term t_i in the collection. Each sample belongs to one of K classes $1, 2, \cdots, K$.

(2) Let $\overline{tf_i} = \frac{1}{N}(tf_{i1} + tf_{i2} + \cdots + tf_{iN})$ be the sample average of these random variables in terms of t_i.

(3) Let $\overline{tf_{ki}} = \sum_{j=1}^{N} tf_{ij}I(d_j, C_k)/N_k, (k = 1, \cdots, K)$ be the sample average of term t_i in category C_k, where $I(d_j, C_k)$ is an indicator to discriminate whether document d_j belongs to C_k, and N_k is the total samples in class k.

According to Lindeberg-Levy central limit theorems (LV CLT) [1], $\overline{tf_i}$ is approximately normal with mean μ_i and variance $\frac{1}{N}\sigma_i^2$, denoted as $\tilde{N}(\mu_i, \frac{1}{N}\sigma_i^2)$; And $\overline{tf_{ki}}$ is approximately normal with mean μ_i and variance $\frac{1}{N_k}\sigma_i^2$, denoted as $\tilde{N}(\mu_i, \frac{1}{N_k}\sigma_i^2)$.

Then we know that $\overline{tf_{ki}} - \overline{tf_i}$ is also approximately normal distributed with mean 0 and variance $(\frac{1}{N_k} - \frac{1}{N})\sigma_i^2$. The variance (Var) is induced as follows:

$$
\begin{aligned}
&Var(\overline{tf_{ki}} - \overline{tf_i}) \\
&= Var\left(\left(\frac{1}{N_k} - \frac{1}{N}\right) \sum_{j \in C_k} tf_{ij} + \frac{1}{N} \sum_{j \notin C_k} tf_{ij}\right) \\
&= \frac{(N - N_k)^2 \times N_k \times \sigma_i^2}{N^2 \times N_k^2} + \frac{(N - N_k) \times \sigma_i^2}{N^2} \\
&= \left(\frac{1}{N_k} - \frac{1}{N}\right) \times \sigma_i^2.
\end{aligned} \tag{1}
$$

Besides, we define the pooled within-class deviation as follows:

$$
s_i^2 = \frac{1}{N - K} \sum_{k=1}^{K} \sum_{j \in C_k} (tf_{ij} - \overline{tf_{ki}})^2 \tag{2}
$$

According to the definition of the t-test [18], we construct the following formula:

$$
t - test(t_i, C_k) = \frac{|\overline{tf_{ki}} - \overline{tf_i}|}{m_k \cdot s_i} \tag{3}
$$

where s_i is standard deviation, and $m_k = \sqrt{\frac{1}{N_k} - \frac{1}{N}}$.

The Eq. 3 is used to measure whether the means of the two normal distributions (i.e., $\overline{tf_{ki}}$ and $\overline{tf_i}$) have the statistically significant difference. The bigger the value of $t - test(t_i, C_k)$ is, the larger the difference of the means is. For some threshold θ, if the $t - test(t_i, C_k) < \theta$, it implies that the averaged frequency of term t_i in the specific category C_k has the same or similar mean with that in the entire corpus; Otherwise,

it implies the averaged frequency of term t_i in the specific category C_k is significantly different from that in the entire corpus, and the term has more discriminating power for the specific category C_k. Compared with the average of term frequency in the entire corpus, the term t_i occurred many or few times in C_k can be considered as the feature of category C_k.

We combine the category-specific scores of a term into two alternate ways:

$$t - test_{avg}(t_i) = \sum_{k=1}^{K} t - test(t_i, C_k) \qquad (4)$$

$$t - test_{max}(t_i) = \max_{k=1}^{K}\{t - test(t_i, C_k)\} \qquad (5)$$

4. EXPERIMENTAL SETUP

4.1 Data Sets

Reuters-21578 [1]: The Reuters corpus is a widely used benchmark collection [4, 5, 20, 19]. According to the ModApte split, we get a collection of 52 categories (9100 documents) after removing unlabeled documents and documents with more than one class label. Reuters-21578 is a very skewed data set. Altogether 319 stop words, punctuation and numbers areremoved. All letters are converted into lowercase, and the word stemming is applied.

20Newsgroup [2]: The Newsgroup is also a widely used benchmark [4, 5, 20], and consists of 19,905 documents, which are uniformly distributed in twenty categories. We randomly divide it into training and test sets by 2:1, and only keep "Subject", "Keyword" and "Content". The stop words list has 823 words, and we filter words containing non-characters. All letters are converted into lowercase and word stemming is applied.

Each document is represented by a vector in the term space, and term weighting is calculated by standard *ltc* [12], and then the vector is normalized to have one unit length.

4.2 Classifiers

In our experiments, we choose three well-established classifiers for the comparison purpose. They are: Support Vector Machines (SVMs) [3], weighted kNN classifier (*k*NN) [20], and classic Centroid-based Classifier (CC) [7]. The SVMs implementation we use is LIBSVM [2] with linear kernels. For *k*NN, we set $k = 10$ [20]. The similarity measure we use is the cosine function.

4.3 Performance Measures

We measure the effectiveness of classifiers in terms of F_1 widely used for TC. For multi-class task, F_1 is estimated in two ways, i.e., the macro-averaged F_1 (macro-F_1) and the micro-averaged F_1 (micro-F_1), as the following:

$$\text{macro-}F_1 = \frac{\sum_{i=1}^{K} F_1(i)}{K}, \qquad (6)$$

$$\text{micro-}F_1 = \frac{2\bar{p}\bar{r}}{\bar{p} + \bar{r}}, \qquad (7)$$

where $F_1(i)$ is the F_1 value of the predicted ith class, and $\bar{p}$ and $\bar{r}$ are the precision and recall values across all classes,

[1]Available on http://ronaldo.cs.tcd.ie/esslli07/sw/step01.tgz
[2]Available on http://kdd.ics.uci.edu/databases/20newsgroup.

respectively. In general, macro-F_1 gives the same weight to all categories. In contrast, micro-F_1 gives the same weight to each instance, which can be dominated by the performance of common or majority categories.

5. RESULTS

Firstly, We show one case study of t-test in real-world corpus. Tables 1 lists the scores of seven different feature selection functions for the selected four terms in category "acq" from the real-life corpus, i.e., Reuters-21578. Based on the literal meaning, the first two terms, i.e., "acquir" and "stake", are closely related to the content of category "acq", while the last two terms, i.e., "payout" and "dividend", belong to other category. However, according to the χ^2, ECE, and TF methods, we wrongly select "acquir" and "dividend" as the features of category "acq", whereas t-test, IG and MI select the features correctly.

Table 1: The feature values of four terms in "acq".

	acquir	stake	payout	dividend
$t - test$	28.053	22.567	3.272	17.796
χ^2	479.482	270.484	131.104	344.045
IG	0.078	0.042	0.009	0.036
MI	1.283	1.126	0.362	0.830
ECE	0.084	0.050	0.028	0.060
TF	749	646	232	903

Then, we show the performance of t-test on two corpora with three classifiers. For Reuters-21578, the number of feature space is all, 17000, 15000, 13000, 11000, 10000, 8000, 6000, 4000, and 2000, respectively, accounting to ten groups of data sets. On 20 Newsgroup corpus, the original feature space reaches up to 210 thousand and we only select less terms as features to save training time. The dimensionality of feature space is all, 2000, 1500, 1000, 500, and 200, respectively, accounting to six groups of data sets.

For χ^2, MI, and t-test methods, we tested the two alternative combinations, i.e., *averaged* and *maximized* ways. We observed that the averaged way was always better than the maximized way for multi-classes problem. Thus we only report the best results of three methods.

5.1 Performance of t-test with kNN classifier

The macro-F_1 and micro-F_1 of five methods with *k*NN on imbalanced Reuters-21578 are shown in Fig. 1, Fig. 2, respectively. It is clear that t-test, χ^2, and ECE achieve evidently better performance than MI and IG in terms of macro-F_1. However, the diversity among the three methods is small. As shown in Fig. 1, when the number of feature space is larger than 13000, χ^2, and ECE is a little better than t-test; However, when the number of features falls in [8000, 13000], t-test performs the best macro-F_1.

The micro-F_1 of five methods increases as the number of features decreases, as shown in Fig. 2. It demonstrates that *k*NN often obtains better performance with less features. Our t-test method performs consistently the best in distinct feature dimensionality, and the highest micro-F_1 of t-test is 89.8% when the number of features is 4000, which improves up to 4.2% than χ^2. IG achieves the worst performance in the all experiments on skewed corpus with *k*NN.

As shown in Fig. 1 and Fig. 2, for unbalanced multi-class tasks, we find IG is inferior to MI in terms of both macro-F_1
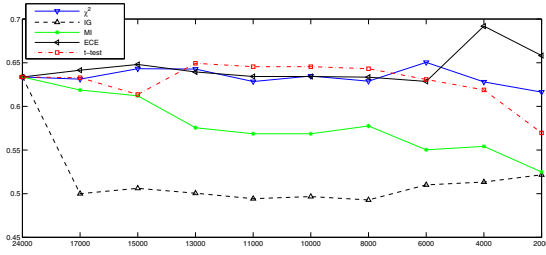
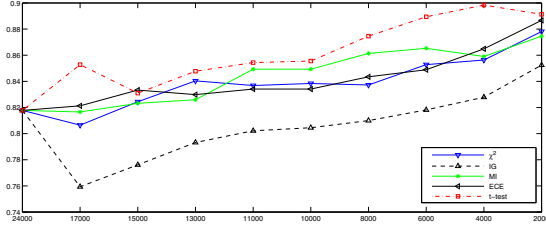**Figure 1: The comparative curves of five methods with kNN on Reuters-21578 in terms of macro-F_1.**

Figure 2: The comparative curves of five methods with kNN on Reuters-21578 in terms of micro-F_1.

and micro-F_1, whereas IG is superior to MI for binary classification tasks according to the comparative experiments of Yang et al [20]. The conflict shows that feature selection methods depends on the practical classification problem.
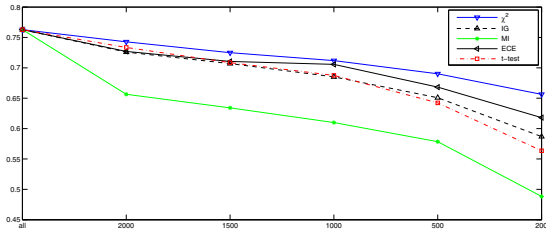

Figure 3: The comparative curves of five methods with kNN on 20 Newsgroup in terms of micro-F_1.

Because macro-F_1 on balanced corpus is close to micro-F_1, we only show the results of micro-F_1 on 20 Newsgroup. As shown in Fig. 3, the micro-F_1 of both χ^2 and IG are slightly better than our t-test method, and the four methods are obviously better than MI. Especially, the performance of IG is comparable to χ^2, and ECE on balanced corpus.

5.2 Performance of t-test with SVMs classifier

Fig. 4 and Fig. 5 depict the macro-F_1 and micro-F_1 of different methods on the Reuters-21578 corpus using SVMs. The t-test, χ^2, and ECE methods perform similar performances, which are better than IG and MI methods. Meanwhile, the macro-F_1 scores of three methods increase as the number of features reduces. It is worth noting that MI does better than other methods when the number of features is in [15,000, 24,411], and then MI falls dramatically.

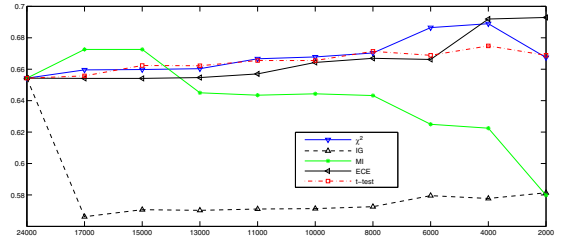The performance of these methods in terms of micro-F_1 on Reuters-21578 corpus is shown in Fig. 5. The micro-F_1

Figure 4: The macro-F_1 of different methods on Reuters-21578 using SVMs.
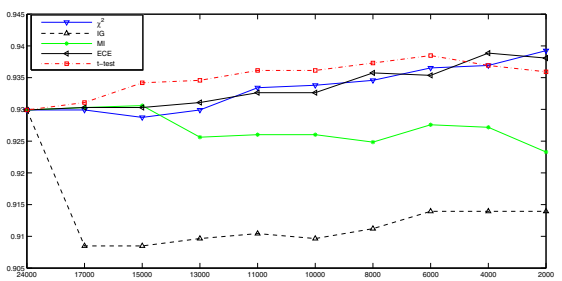

Figure 5: The micro-F_1 of different methods on Reuters-21578 using SVMs.

points of different feature selection methods show a tendency to increase as the number of the features decreases. However, these methods show consistent performance in micro-F_1, and the t-test method is still the best among these methods.
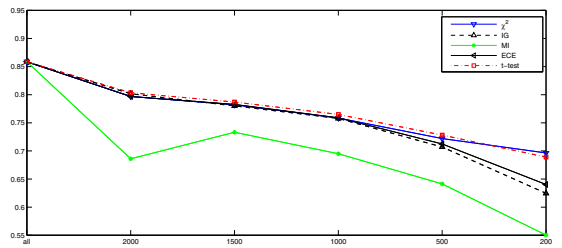

Figure 6: The micro-F_1 of different methods on 20 Newsgroup using SVMs.

Fig. 6 depicts the micro-F_1 of different methods on the 20 Newsgroups using SVM. The trends of the curves are similar to those in Fig. 3. The t-test, χ^2, IG, and ECE achieve similar performances, which are better than MI. Our t-test is slightly better than others.

5.3 Performance of t-test with Centroid-based classifier

For centroid-based classifier, the macro-F_1 of five methods is shown in Fig. 7. We can observe that χ^2, ECE, and t-test do better than MI and IG methods, and χ^2 is slightly better than ECE and t-test. The same conclusion can be done in terms of micro-F_1, as shown in Fig. 8.

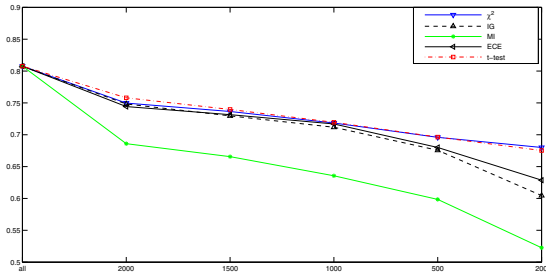Meanwhile, our t-test is slightly better than χ^2, ECE, and
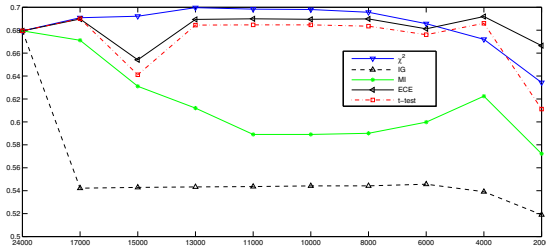
Figure 7: The macro-F_1 of five methods on Reuters-21578 using centroid-based classifier.
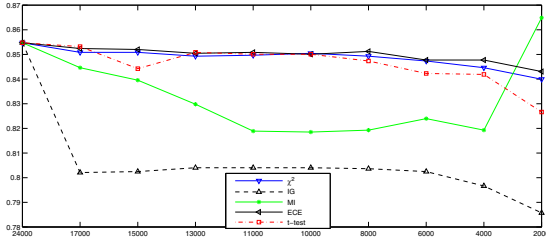
Figure 8: The micro-F_1 of five methods on Reuters-21578 using centroid-based classifier.

IG methods on 20 Newsgroup corpus. The four methods outperform the MI method significantly.

Figure 9: The micro-F_1 of five methods on 20 Newsgroup using centroid-based classifier.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new feature selection method based on term frequency and t-test. Then we compare our approach with the state-of-the-art methods on two corpora using three classifiers in terms of macro-F_1 and micro-F_1. Extensive experiments have indicated that our new approach offers comparable performance with χ^2, and ECE, even slightly better than them. In future work, we will verify our method on more text collections.

7. REFERENCES

[1] P. Billingsley. *Probability and Measure (Third ed.)*. John Wiley & sons, 1995, 357-363.

[2] C. Chang and C. Lin. Libsvm: a library for support vector machines. 2001.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 1995, (20), 273-297.

[4] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 1993, 19(1), 61-74.

[5] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 2003, 3, 1289-1305.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003, 3, 1157-1182.

[7] E.-H. Han and G. Karypis. Centroid-based document classification: Analysis & experimental results. *In: Proceedings of PKDD*, 2000.

[8] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. *In: Proceedings of ICML*, 1997, 170-178.

[9] S. Li, R. Xia, C. Zong, and C. Huang. A framework of feature selection methods for text categorization. *In: Proceedings of 47th ACL and the 4th AFNLP*, 2009.

[10] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop*, 1998.

[11] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. *In: Proceedings of ICML*, 1999.

[12] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, 24(5), 513-523.

[13] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput Surv*, 2002, 34(1), 1-47.

[14] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.*, 2002, 99: 6567-6572.

[15] H. Uguz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowl.-Based Syst.*, 2011, 24(7): 1024-1032.

[16] A. Unler, A. Murat, and R. B. Chinnam. mr2pso: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Inf. Sci.*, 2011, 181(20):4625-4641.

[17] Y.-Q. Wei, P.-Y. Liu, and Z.-F. Zhu. A feature selection method based on improved tfidf. *In: Proceedings of the ICPCA*, 2008, 94-97.

[18] S. William. The probable error of a mean. *Biometrika*, 1908, 6(1), 1-25.

[19] S.-M. Yang, X. Wu, and Z. Deng. Relative term-frequency based feature selection for text categorization. *In: Proceedings of ICMLC*, 2002.

[20] Y.-M. Yang and J.-P. Pedersen. A comparative study on feature selection in text categorization. *In: Proceedings of ICML*, 1997, 412-420.

[21] N.-N. Zhou and L.-P. Wang. A modified t-test feature selection method and its application on the hapmap genotype data. *Geno. Prot. Bioinfo.*, 2007, 5(3-4), 242-249.