An incremental learning algorithm considering texts' reliability

Xinghua Fan, Shaozhu Wang
College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing, China

***Abstract*—The sequence of texts selected obviously influences the accuracy of classification. Some sequences may make the performance of classification poor. For overcoming this problem, an incremental learning algorithm considering texts' reliability, which finds reliable texts and selects them preferentially, is proposed in this paper. To find reliable texts, it uses two evaluation methods of FEM and SEM, which are proposed according to the text distribution of unlabeled texts. The results of the last experiments not only verify the effectiveness of the two evaluation methods but also indicate that the proposed incremental learning algorithm has advantages of fast training speed, high accuracy of classification, and steady performance.**

Keywords-text classification; incremental learning; reliability; text distribution; evaluation.

I. INTRODUCTION

Conventional methods of text classification, for example, Centroid, Native Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and so on, which are not incremental learning methods, obtain the texts' classification model according to existing labeled training set. However the training set can't be obtained in one time; the methods above are not always effective.

Incremental learning can solve the above problem very well. With the advancement of classification process, in the incremental learning, the scale of training set expands unceasingly; new texts are labeled and added to the training set gradually. Among those text candidates, which texts to select first is the critical point of this classification.

There are two models of selecting texts to add into labeled training set: passive classification model and active classification model.

Passive classification model, which selects the training texts randomly and accepts the text information passively; it believes that the training texts distribute independently in most of classification learning, so passive classification model has obvious deficiency:

- Make the noise spread down, affect the accuracy of classification.
- Ignore the relationship among texts in new incremental training set.

Active classification model selects texts actively. It is a subconscious and higher level learning model, which selects the optimized texts to improve classifier's performance. So compared with passive learning, active learning attracts more researchers' attentions. Reference [1] proposed an algorithm to select a text by calculating the 0-1 loss rate every time, and the algorithm improved the performance of classifier. But large amount of calculation and high time complexity are the algorithm's shortages. Reference [2] proposed an algorithm to select some texts by clustering. This algorithm reduced the training time, but it would be affected by noise data easily and lead to large fluctuations of classifier's performance. No matter the algorithm of selecting one text or that of batch selection [1][2][3][4][5], texts are selected by external evaluation algorithms which need a lot of additional computing, so most of incremental learning algorithms have poor efficiency.

From above, the method to select texts is very important. A good method not only improves the classifier's performance but also reduces the training time. For solving this problem, an incremental learning algorithm considering texts' reliability is proposed in this paper. It includes two evaluation methods named first evaluation method (FEM) and second evaluation method (SEM), which select new texts according to the results in Reference [6], are proposed in this paper. Reference [6] showed that classifier's performance will be improved obviously when the correctly labeled texts are added preferentially. And these two methods are complementary to each other and have low computational complexity, which make full use of useful information among texts and the intermediate data-out in the process of training classifier. For incremental bayesian model [1] can make good use of its prior knowable, it is used to improve the availability of the algorithm proposed. The structure of this paper is organized as follows: the algorithm is introduced in detail in Section II. Section III demonstrates experimental results on artificial and real datasets. We conclude our study in Section IV.

II. AN INCREAMTAL ALGORITHM CONSIDERING TEXTS' RELIABILITY

In this section, a new incremental algorithm will be introduced in detail. The two FEM and SEM methods are important parts of the algorithm .They are inspired from the regularity of texts' distribution, so the corresponding regularity of texts' distribution will be introduced first, and then introduce evaluation methods and their relation. The details of each step of the new algorithm will be given in the end of this section.

A. The first evaluation method (FEM)

Given text vector $d = (W_1, W_2, \ldots, W_n)$ ($W_i = 0$ or 1). If the i-th feature appear in the text, $W_i = 1$, otherwise $W_i = 0$. Supposed that $p_{ki} = \{W_k = 1 | c_i\}$, and $p\{\bullet\}$ is the probability for incident $\{\bullet\}$.The discriminant function[7] of Naive Bayesian classifier can be expressed as:

$$c^* = \arg\max(\log P\{c_i\} + \sum_{k=1}^{D} \log(1 - p_{ki}) + \sum_{k=1}^{D} W_k \log \frac{p_{ki}}{1 - p_{ki}}) \quad (1)$$

Supposed that:

$$MV_i = \log P\{c_i\} + \sum_{k=1}^{M} \log(1 - p_{ki}) + \sum_{k=1}^{M} W_k \log \frac{p_{ki}}{1 - p_{ki}}) \quad (2)$$

$$MV_{\max} = \max_{c_i \in C}(MV_i) \quad (3)$$

$$MV_{\sec} = \underset{c_i \in C}{\sec ond}\ (MV_i) \quad (4)$$

MV_i is the probability of text vector d, which is estimated by feature and belongs to $c_i \in C$, and C is the predefined type set. $MV_{\max}$ is the maximum of all probabilities in text vector d; $MV_{\sec}$ is the second maximum of all probabilities in text vector d.

The value of rewritten MV_i is negative, normalizing for MV_i:

$$p = MV_{\max} / MV_{\sec} \quad (5)$$

Take the corpus, which will be introduced in section III, as samples. We randomly divide the 6000 texts into 3 groups of datasets. Each group contains a labeled training set of different scales which are 20 texts, 200 texts, 2000 texts, and a common new incremental training set composed of 400 unlabeled texts. Then construct the classifier and classify the new incremental training set. The relationship between the p-value and the number of misclassified texts is shown in table I.

The largest set of the correct texts refers to the texts contained within the p-value, where the misclassified text appears for the first time. Table I shows that the misclassified texts appear and increase gradually with the p-value changing. The greater the p-value is, the more misclassified texts appear. If a set within p-value contains no misclassified texts, it is the correct interval, and names the set of the others texts as fuzzy interval. Table I plus table II, show that with the size of labeled set increasing, more and more texts are distributed in the correct interval. In addition, table I plus table II, show the existence of the correct interval has nothing to do with the scale of labeled texts; the scale only affects the number of texts in correct interval.

TABLE I. THE RELATIONSHIP BETWEEN P-VALUE AND THE NUMBER OF MISCLASSIFIED TEXTS

p's range	**The number of misclassified texts**		
	Labeled texts(20)	***Labeled texts (200)***	***Labeled texts (2000)***
(0,0.5)	0	0	0
[0.5,0.6)	3	0	0
[0.6,0.7)	4	1	1
[0.7,0.8)	6	1	1
[0.8,0.9)	12	7	3
[0.9,1]	22	6	5

TABLE II. THE RELATIONSHIP BETWEEN LABELED TEXTS' SCALE AND PERCENTAGE OF THE LARGEST SET OF THE CORRECT TEXTS

Labeled texts	20	200	2000
Percentage (%)	40.25	78.75	79.25

Table II shows that when the number of labeled texts is equal or more than 200, nearly 80% of the texts are distributed in the correct interval. As the initial labeled texts are few, in order to maximize the number of the new incremental unlabeled texts falling into the correct interval, the new incremental training set is divided into a number of subsets each containing 100 texts. Carrying out incremental learning among the subset takes advantage of the size and performance of intermediate classifiers.

From the regularity of texts' distribution mentioned, the method of FEM is proposed as follows:

FEM: in the output of classifier, if p which is calculated by formula $p = MV_{\max} / MV_{\sec}$ not exceeds a threshold α, corresponding texts are all corrected classified texts.

In order to determine the value of α, take the corpus, which will be introduced in section III, as samples. Take 5 labeled texts each category to construct training set with 20 labeled texts, and classify for 600 new texts by constructed initial classifier, the relationship between the *p-value* and the distribution of misclassified texts is shown in Fig. 1.

Fig. 1 shows that the value of α should be between 0.5 and 0.6, in order to ensure that the texts in this interval are all up to the requirements, α 's value should be set to 0.5.

B. The second evaluation method (SEM)

After the FEM assessment, the texts incorrectly labeled by the current classifier concentratedly distribute in fuzzy interval. Deal the texts in fuzzy interval with Affinity Propagation (AP) clustering [8], and get many clusters. In each cluster, the first text is a representative for the others. And most of the texts have the same label as the first text in each cluster. The results of the experiments in Reference [3], which only uses noun as features, show that: more than 90% of the texts have the same label as the representative text. So the result can be used for judging whether the classifier is able to correctly identify the cluster. Take the corpus, which will be introduced in section III, as samples. We randomly get a group of dataset from the 6000 texts.

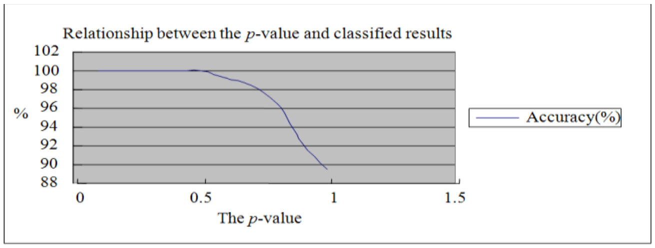

Figure 1. Relationship between the p-value and distribution of misclassified texts

TABLE III. THE ACTUAL LABELS AND OBTAINED LABELS OF CLUSTERS

Texts' actual category	Texts' category of current classifier
2 2 2	2 2 2
1 1 3 1 1 1	1 1 3 1 2 1
3 3 3 3 4 3	3 3 2 1 3 2
4 1 4	4 1 4
3 2 3 3	3 1 2 3

The dataset contains a labeled training set composed of 5 texts each category and a new incremental training set composed of 600 texts. Classify for 600 new texts by initial classifier constructed, α 's value is set to 0.5, do AP clustering for texts in fuzzy intervals. Analyzing the first 5 clusters, their actual labels and obtained labels are shown in table III.

Analyze the label of the third cluster, a conclusion is got, the labeled training set will be introduced four incorrectly labeled texts by the current classifier. In order to avoid this, we only join the texts which have the same label as the representative text into labeled training set, compute the ratio $\beta = num1/num2$, where $num1$ is the number of the texts which have the same label as representative text, $num2$ is the number of the whole cluster. Set a threshold δ , and it means that the current classifier can't correctly identify the cluster if β is less than δ , remove the cluster. And put forward the method of SEM as follows:

SEM: Classify the texts in each cluster by the current classifier, and then calculate the ratio $\beta = num1/num2$. Set a threshold δ **,** if β is not less than δ , it believes that the texts in corresponding cluster can be identified by the current classifier.

In order to determine the value of parameter δ , take the corpus, which will be introduced in section III, as samples. Take 5 texts as labeled texts each category to construct training set with 20 labeled texts, and classify for 600 new texts by initial classifier constructed, the fuzzy intervals are obtained when α =0.5, the relationship between the value of β and learning results of texts in the fuzzy intervals is shown in Fig. 2. As is shown in Fig. 2, the learning performance of classifier is the best when the value of δ near 0.8.

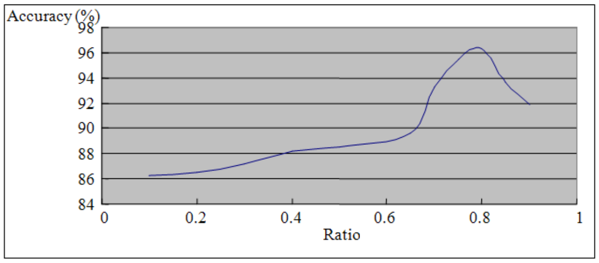

Figure 2. Relationship between the value of β and accuracy of texts' learning

TABLE IV. MACRO_F OF LEARNING FOR TEXTS IN FUZZY INTERVALS

Scale of labeled set	20	200	2000
Macro_F (%)	79.68	82.36	86.72

TABLE V. MACRO_F OF LEARNING BY SEM ONLY

Scale of labeled set	20	200	2000
Macro_F（%）	91.04	94.57	99.18

C. Complementarities of FEM and SEM

After the FEM assessment, if continue to do incremental learning for texts in fuzzy intervals by current classifier, the accuracy of learning is not very good. Take the corpus, which will be introduced in section III, as samples. Take 5 texts as labeled texts each category to construct labeled training set with 20 labeled texts, and classify for 600 new texts by constructed initial classifier, do incremental learning sequentially for texts in fuzzy intervals when α is equal to 0.5, and take 50 and 500 labeled texts as well. Accuracy is shown in table IV.

As is shown in table IV, with the expansion of labeled training set's scale, the accuracy is better. And combined with figure 2, the accuracy rises by 79.68% to more than 96% by adding SEM.

Current classifier's performance need to be considered in SEM, so it will obtain better results when knowledge of classifier is abundant. If performance of initial classifier is not very good, it will yield big error in calculating the value of β , noise data is introduced and finally lead to the bad performance of classifier. Set the value of δ to 0.8, the results of incremental learning by SEM only (use the same corpus with table IV) are shown in table V.

As is shown in table V, if evaluated by SEM only, the final classifier's performance is obviously affected by initial classifier. The reason is that noise data is introduced into labeled training set in previous iteration. As is shown, the larger scale the initial labeled training set is, the better result the SEM can obtain. And eighty percent of texts are in the correct interval after evaluating by FEM which can lead to obtain a large amount of labeled training set. So FEM and SEM are complementary to each other.

D. Description of algorithm

The two mentioned evaluation methods provide theory basis for the new algorithm proposed in the paper. The algorithm, which uses the two evaluation methods to make the reliable texts join labeled set preferentially, improves the performance of the classifier and reduces the influence by noise data. Because the proportion of texts in correct interval is influenced by the scale of the initial labeled set, divide the unlabeled set into some subsets. So more texts can be in correct interval by intermediate classifies. The algorithm can be described concretely as follows:

Input: Labeled training set $D = \{d_1, d_2, \ldots, d_N\}$

New incremental training set $T = \{t_1, t_2, \ldots, t_m\}$

Output: Classifier *C*

Step1: Use the CHI formula to do the feature selection for training set *D*, and learn a classifier;

Step2: If $T = \Phi$ (Φ is the empty set), go to step5;

Step3: Randomly select 100 texts from *T*, classify each text t_p in new incremental training set *T* by current classifier *C*, select correct texts estimated by FEM to form a new subset $T' \subset T$, and add them into the training set *D*, the rest is added into the untrusting set *U*;

Step4: $T = T - T'$, go to step1;

Step5: If $U = \Phi$, return the classifier, and end the algorithm; else continue;

Step6: Do clustering for the untrusting set *U*, formed *k* subsets $U = \{R_1, R_2, \ldots, R_k\}$, remove the subsets which only have a single text to set U , then select the first text of each cluster respectively to construct a representative text set $r = \{r_1, r_2, \ldots, r_m\}$, $m < k$;

Step7: If $r = \Phi$, go to step5, else for each of the text $r_i \in r$, to repeat the follows:

a) Classify texts r_i by current classifier C , and obtain the label C_p ;

b) Classify other texts in subset Ri which ri is in by current classifier C , and calculating the ratio (β) of num to NUM, where NUM is the total number of texts in the cluster which ri is in and num is the number of texts which are classified the same category with ri;

c) If $\beta > \delta$, join the texts including ri in Ri, which are classified the same label with ri, into T'' , then update the set $r = r - r_i$;

d) $D = D + T''$, $U = U - T''$, use the CHI formula to select features for training set D, and learn classifier C .

III. Experiments

Five experiments are designed in this paper:

Exp.1: Verify the effectiveness of the correct set division.

Exp.2: Verify the effectiveness of fuzzy data processing.

Exp.3: Verify the effectiveness of subset division.

Exp.4: Verify the high efficiency and steady performance of the proposed method.

Exp.5: A test of training time and learning performance of different scales of new incremental training set.

A. The datasets of experiments

Datasets: The datasets used in experiments are all from netease and sina, which including four categories, and have total 6000 Chinese texts. In the 6000 Chinese texts, category of Olympics, Buddhism, Military and Computer has 1500 texts respectively. Form eight groups of corpus used in Exp.1, Exp.2, Exp.3 and Exp.4. Each group contains 5 initial labeled texts and 100 unlabeled texts each category from the 6000 texts randomly. And form four groups of corpus used in Exp.5. Each group contains a training set with 5 labeled texts each category, and a new incremental training of different scales which are 400 unlabeled texts, 800 unlabeled texts, 1200 unlabeled texts. The same texts mustn't appear in both initial labeled training set and unlabeled training set.

B. The feature selection in experiments

The feature selection method of CHI is used in experiments:

$$\chi^2(w,c) = \frac{N(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (6)$$

Where, *c* is the category, *w* is the feature, *N* is the number of texts, *A* is times of *w* and *c* both appeared, *B* is times of *w* appeared but *c* not appeared, *C* is times of *c* appeared but *w* not appeared. *D* is times of *w* and *c* both not appeared.

C. Performance's assessment

Precision: $P = \frac{N_1}{N_2} \times 100\%$

Recall: $R = \frac{N_1}{N_3} \times 100\%$

Macro average: $\text{Macro\_F} = \frac{2 \times P \times R}{P + R} \times 100\%$

Where, N_1 is the number of texts correctly classified in a category, N_2 is the number of texts classified in a category, N_3 is the number of texts in a category of test set.

D. Experimental Results

1) The methods in experiments are defined as:

NBTS: Incremental method considering texts' reliability proposed in this paper.

NBSS: Incremental method with SEM.

NBFS: Incremental method with FEM.

NBS: Incremental method without division subset.

NBKC: Quick clustering based incremental method proposed in reference [4].

EM: The standard Expectation Maximization (EM) algorithm [9].

2) The parameters setting

From the second section, if the classifier's performance is the best, the parameter α is equal to 0.5 and δ is equal to 0.8.

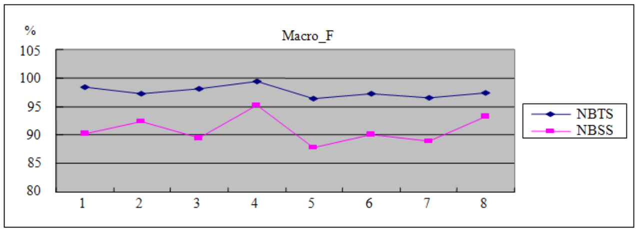

Figure 3. The learning results of NBTS and NBSS

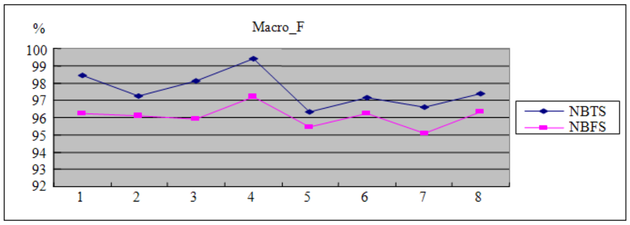

Figure 4. The learning results of NBTS and NBFS

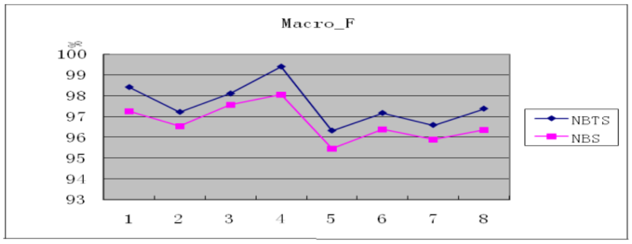

Figure 5. The learning results of NBTS and NBS

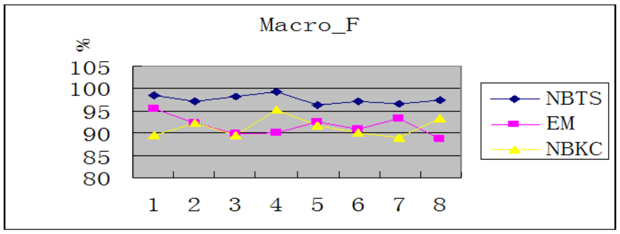

Figure 6. The learning results of the mentioned three incremental methods

TABLE VI. THE AVERAGE TIME CONSUMING OF THE MENTIONED TWO INCREMENTAL METHODS IN EXP. 3

Method	Average time consuming(s)
NBTS	115
NBS	135

TABLE VII. THE AVERAGE TIME CONSUMING OF THE MENTIONED THREE INCREMENTAL METHODS IN EXP. 4

Method	Average time consuming(s)
NBTS	115
EM	200
NBKC	1865

TABLE VIII. THE TRAINING TIME IN DIFFERENT SCALES OF NEW INCREMENTAL TRAINING SET MENTIONED IN THIS PAPER

Group number	The scale of new incremental training set and its training time(s)					
	400	*Time(s)*	*800*	*Time(s)*	*1200*	*Time(s)*
1	98.42	121	96.93	203	97.38	298
2	97.22	94	98.39	215	98.97	307
3	98.11	106	97.74	198	97.12	287
4	99.41	131	98.96	234	96.59	279

3) The results of experiments

Results of Exp. 1-Exp. 4 are shown in Fig. 3-Fig. 6 respectively.

The average time consuming of the methods in Exp. 3 and Exp. 4 are shown in table VI and VII respectively.

Results of Exp.5 are shown in table VIII.

E. Analyses of the experimental results

- Exp.1 shows that the classifier's performance is greatly improved by adding the correctly classified texts to labeled training set, Macro_F increases by about 7% relative to use SEM only. FEM's effectiveness is verified.
- Exp.2 shows that after using SEM to deal with fuzzy data, the classifier's performance increases by 2%. SEM's effectiveness is verified.
- Exp.3 shows that the learning method with division subsets not only improves the classifier's performance, but also shorts the train time. With increase of labeled training set's scale, more and more unlabeled texts lie in the correct interval. The intermediate classifiers are fully used by dividing subsets, more texts are added by FEM, the performance of the classifier is improved, the number of texts in fuzzy interval is reduced and clustering and text selection's time is shorter.
- Exp.4 shows that the classifier trained by proposed algorithm has better and steadier performance, for it decreases the disturbance of noise in the data sets.
- Exp.4 and Exp.5 show that the classifier trained by proposed algorithm has better performance and shorter train time than classifiers trained by other algorithms. The algorithm is more suitable for dealing large data.

IV. CONCLUSIONS

An incremental learning algorithm considering texts' reliability is proposed in this paper. Firstly, the new incremental training set is divided into subsets and the FEM method is used to find out the correct set interval of the subset, which made the number of labeled training set greatly increase.

Then the remaining fuzzy data was dealt by AP classification, and the learning sequence of noise data is further dealt by SEM. The experimental results show that the proposed algorithm is less affected by noise data and the performance of classifier is relatively stable. And the proposed incremental learning algorithm can train a classifier quickly.

ACKNOWLEDGMENT

The research is supported in part by the National Natural Science Foundation of China under grant number 60703010, cultivate project for significant research plan of the National Natural Science Foundation of China under the grant number 90924029 named "Research on the mechanism and related technology of public opinion on internet for abnormal emergency" and the Nature Science Foundation of Chongqing province in China under grant number CSTC, 2009BB2079.

REFERENCES

[1] Xiujun Gong, Shaohui Lin and Zhongzhi Shi, "An Incremental Bayes Classification Model," Chinese Journal of Computers, vol. 25, no. 6, pp. 645-650, 2002.

[2] Houfeng Ma and Xinghua Fan. "Improved incremental Bayes classification algorithm," Chinese Journal of Scientific Instrument, vol.28, no. 8III, pp. 312-316,2007.

[3] Houfeng Ma, Xinghua Fan and Ji Chen, "An incremental Chinese text classification algorithm based on quick clustering," 2008 International Symposiums on Information Processing. IEEE Computer Society, 2008, pp. 308-312.

[4] Xinghua Fan, Zhiyi Guo and Houfeng Ma, "An improved EM-based Semi-supervised Learning Method. 2009 International Joint Conference on Bioinformatics," Systems Biology and Intelligent Computing (IJCBS'09). IEEE Computer Society, 2009, pp. 529-532.

[5] Yimin Wen, Yang Yang and Baoliang Lv, "Study on Ensemble Learning and Its Application in Incremental Learning," Journal of Computer Research and Development, vol. 42 (supplement), pp. 222-227,2005.

[6] M. Li, Z. H. Zhou. SETRED: Self-training with editing[C]. In Proceedings of 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05), Hanoi, Vietnam, LNAI 3518, 2005, pp. 611-621.

[7] X. H. Fan and M. S. Sun, "A High Performance Two-Class Chinese Text Categorization Method," Chinese Journal of Computers, vol. 29, no.1, pp. 124-131, 2006.

[8] Brendan J. Frey and Delbert Dueck, "Clustering by Passing Messages Between Data Points", Science, vol.315, February 2007, pp. 972-976.

[9] Arthur Dempster, Nan Laird and Donald Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, vol.39, no.1, pp.1–38, 1977.

[10] Y. Jiang and Z. H. Zhou, "Editing training data for KNN classifiers with neutral network ensemble," In Proceeding of the 1st International Symposium on Neural Networks, Dalian, China, 2004, pp. 356-361.

[11] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[12] Jinshu Su, Bofeng Zhang and Xin Xu, "Advances in Machine Learning Based Text Categorization," Journal of Software, vol. 17, no. 9, pp. 1848-1859, 2006.

[13] A. McCallum and K. Nigam, "A comparison of event models for Naïve Bayes text Classification," AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41-48.

[14] D. D. Lewis, R. E. Schapire, J. P. Callan and R. Papka, "Training algoirthms for linear text classifiers," in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 298-306.

[15] E. H. Han, G. Karypis and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," University of Minnesota, 1999, pp. 11-12.

[16] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Machine Learning:ECML-98, Tenth European Conference on Machine Learning, 1998, pp. 137-142.

[17] M. Ruiz and P. Srinivasan, "Hierarchical text categorization using neural networks," Information Retrieval, vol. 5, no. 1, pp. 87-118, 2002.

[18] Z. H. Zhou and W. Tang, "Selective ensemble of decision trees," Lecture Notes in Artificial Intelligence 2639, Berlin: Springer, 2003, pp.476-483.

AUTHORS PROFILE

Xinghua Fan was born in 1972, in Chongqing, China. He received his Ph. D. from Chongqing University. He is a professor in Chongqing University of Posts and Telecomunications. His research interests include artificial intelligence, natural language processing and information retrieval.

Shaozhu Wang is a postgraduate in Chongqing University of Posts and Telecomunications. She was born in 1986, in Fuzhou, Fujian province, China. Her research interests include Chinese information processing and text classification.