International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.12 (2015), pp.195-206
http://dx.doi.org/10.14257/ijmue.2015.10.12.20

Short Text Classification Algorithm Based on Semi-Supervised Learning and SVM

Chunyong Yin[1], Jun Xiang[1], Hui Zhang[1], Zhichao Yin[2] and Jin Wang[1]

[1]*Jiangsu Key Laboratory of Meteorological Observation and Information Processing, School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*Nanjing No.1 Middle School, Nanjing, Jiangsu, Postal code 210001, China*

Abstract

Short text is a popular text form, which is widely used in real-time network news, short commentary, micro-blog and many other fields. With the development of the application such as QQ, mobile phone text messages and movie websites, the size of data is also becoming larger and larger. Most data is useless for us while other data is significant for us. Therefore, it is necessary for us to extract the useful short text from the big data. However, there are many problems with the short text classification, such as fewer features, irregularity and so on. To solve these problems, we should pretreat the short text set first, and then choose the significant features. This paper use semi-supervised learning method and SVM classifier to improve the traditional methods and it can classify a large number of short texts to mining the useful massage from the short text. The experimental results in this paper also show a good promotion.

Keywords: *semi-supervised learning, short text classification, SVM, pretreatment, feature selection*

1. Introduction

Recent years, with the development of the Internet and mobile communication, short text has gradually become a new form of the text, and it is widely used in mobile phone, short message, micro-blog, network comments, and video barrage interaction. Short text has the following features.

(1) Sparsity

The length of each short text message is short, often less than 150 words, maybe it is a few sentences, or even a few words. So it contains very little effective information, which results in the sparsity of short text, and the dimension of feature set is very high. It is difficult to extract accurate and key sample features for classification learning.

(2) Real-time

Information in the form of short text messages on the Internet, most of them update in real time, refresh speed is very fast, like chat information, micro-blog information, comments, *etc.*, and constantly update in seconds, it is difficult to collect. What's more, this part of the dynamic text is very large, which requires that the short text information classification must have a higher efficiency.

(3) Irregularity

Information terminology of short text is not standardized, it includes more popular vocabulary, and results in a lot of noise interference. The noise interference should be removed, or it will affect the classification results.

ISSN: 1975-0080 IJMUE

In the face of large-scale short text, it is difficult to obtain the required information quickly and accurately. So short text classification technology is used to solve the problem and it has extensive application prospects in mining user's interests, hot topic tracking, the analysis of buzzwords and so on. Short text classification technology has become a hot research topic nowadays.

At present, there are many traditional text classification methods like Naive Bayes (NB), Support Vector Machines (SVM), Neural Network (NN), Decision Tree (DT), k-Nearest Neighbor (KNN), but those methods put long text as a research object. To short text, those methods have a poor performance because of the less feature, irregularity and big data. Those traditional text classification methods cannot be simply used to solve the short text classification and it has low efficiency when the size of data is large. So a new method should be designed to solve the problem. This paper use semi-supervised learning and SVM to improve the traditional method and it can classify a large number of short texts to mining the useful massage from the short text.

2. Related Research Statuses

Vapnik [1] proposed Support Vector Machines (SVM) based on the research of statistical learning theory at 1995. SVM had better performance than other methods at that time. It can solve the problems of small samples, non-linear and high dimension and had been used successfully in many fields like text classification, face recognition, fingerprint recognition, handwriting. Gu Bin [2] *et al*. proposed a special procedure called initial adjustments, the procedure adjusts the weights of v-SVC based on the Karush–Kuhn–Tucker (KKT) conditions to prepare an initial solution for the incremental learning. Victor S. Sheng [3] *et al*. presents a modified SVOR formulation based on a sum-of-margins strategy. The algorithm can handle a quadratic formulation with multiple constraints, where each constraint is constituted of an equality and an inequality.

Cover and Hart [4] proposed k-Nearest Neighbor(KNN) algorithm at 1968, it was a common machine learning algorithms, its basic idea is as follows: If a target sample has k most similar samples in the feature space and the k samples belong to a class, then the target sample also belongs to the class. The key part of the KNN algorithm is the similarity calculation method and the selection of k value. At present, there are several similarity calculation methods such as Euclidean distance, the inner product formula and cosine formula and so on.

Maron proposed Naive Bayes (NB) algorithm based on Bayes theory and Probability research and it is researched widely in the field of machine learning [5]. The requirements of two conditions when using this method to classify:

(1) The probability distribution of each category is known;

(2) The number of categories to be made is certain. NB algorithm can be applied to a big database simply and efficiently.

The algorithms above have a good performance on the long text classification but a poor performance on irregular short text. Reference [6] and reference [7] pointed out that we could use online encyclopedia, which had a wide coverage areas of knowledge to classify and cluster the short text and it could expand the short text massage, but the method did not take learning problem of lacking labeled samples and the values of unlabeled training samples into consideration. Shi wei [8] built fuzzy emotion ontology based on the sentiment analysis of micro-blogs to classify those micro-blogs by emotion efficiently.

3. Short Text Classifications

3.1. Pretreatment of Short Text

There is a lot of useless information in most short text, which would reduce the effect of text classification, so we need to pretreat the collected corpus.

It is easy to distinguish each word in English because of the blank space between two words. There is no blank space between Chinese words, so we must do word segmentation operation on the Chinese short text firstly. It is a measure to use strings to match participles in the built database, if matching successfully those characters will be seen as a single word. There are some common match methods including positive (negative) maximum matching method and optimal matching method [9]. From the word segmentation operation we can see that some words appear frequently without useful information, this kind of words is called 'stop-words', such as "because", "so", "although", "but" and so on. Those stop-words should be deleted to ensure the classification effect.

This paper used the existing invalid data dictionary information Z to fuzzy match the word and deleted the useless information. However, there was still a lot of unlabeled useless information in the short text, so in this paper we used a semi-supervised learning method to label the unlabeled information in an iterative way to increase the content of the data dictionary.

3.2. Feature Expression

A particular representation form of words in English or Chinese is used for computer to identify them. This paper used Vector Space Model (VSM) to express short text. VSM was proposed by G.Salton [10]. The smallest data unit in the model can be operated in the form of feature item.

For example, if a text T is seen as a n-dimensional vector in the vector space. It shows as follows:

$$T=((t_1, w_1), (t_2, w_2), \cdots, (t_n, w_n))$$

Among them, t is the feature item of the text T, w is the weight value to the feature item. With the greater w, the more information feature item includes and lower w, less information [11].

3.3. Feature Selecting Methods

The classification results are closely related with the feature selecting methods [12-13]. An efficient feature selecting method can not only reduce the feature dimension but also delete the useless feature. A typical feature selecting method includes four following processes: subset generating, subset evaluation, termination conditions and result validation [14-15].

This paper uses a feature selecting method called chi-square test, the chi-square test is also called χ^2 test, and it is a common feature selecting method. Its basic idea is that assuming that there is a first-order χ^2 distribution of DOF (degrees of freedom) between feature item and target categories. The method could calculate degree of correlation to judge whether fit the feature item is to the classification. The correlation between feature item and target categories is proportional to the value of χ^2 test. Greater the value is, greater the relevance is between feature item and target categories and it would have high probability to be selected into feature subset. The formula of chi-square test is as follows.

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} p(c_i)\chi^2(t, c_i)$$

Among them, *A* is the text number including feature item *t* and belonging to categories c_i. *B* is the text number including feature item *t* but not belonging to categories c_i. *C* is the text number belonging to categories c_i but not including feature item *t*. *D* is the text number neither including feature item *t* nor belonging to categories c_i. *N* is the total number of all texts, *A*+*B*+*C*+*D*=*N*.

There are two calculation methods to calculate χ^2 test value of some feature item *t*. The formulae are as follows:

$$\chi^2_{\max}(t) = \max_{1 \le i \le m} \{\chi^2(t, c_i)\}$$

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} p(c_i)\chi^2(t, c_i)$$

3.4. Feature Weight Calculation

Using the vector space model representation of the text, each feature item represents its degree of contribution to the text. If a feature item in the text contains a large amount of information that represents the text, then the feature item of the degree of contribution to the text classification is larger. Otherwise, the feature item of the degree of contribution to the text classification smaller.

How to determine how much text information is included in the feature item is determined by the weight of the feature item. The higher the weight is, the more text information is contained in the feature item. Otherwise, it shows that the feature item contains less text information. Therefore, the calculated results of the feature weights will directly affect the classification results. Nowadays several weight calculation methods are usually used in the researches, such as Boolean function, word frequency function, TF-IDF function and so on. The following paragraphs are a brief introduction of the word frequency function and TF-IDF function.

1) Word frequency function

Word frequency weighting function is a simple, intuitive weight calculation method. It represents the number of times a feature item appears in the text. The more often a feature item appears in the text, the more important it is. The following is the word frequency function's calculation formula:

$$w_{td} = tf_{td}$$

Among them, w_{td} represents the times of feature *t* appears in text *d*, and the weight of feature *t*.

2) TF-IDF function

TF-IDF weight is widely used in text classification for feature weight calculation, the main idea is that if a feature item's word frequency is high in a text, and rarely appears in the other text, then we can consider that this feature item has a good ability to distinguish categories, suitable for classifying. TF-IDF is calculated according to the proportional to a feature item's word frequency in the text, and inversely proportional to the number of texts that contain the feature item in the text.

$$w_{td} = tf_{td} \times idf_t = tf_{td} \times \log\left(\frac{N}{n_t} + 0.5\right)$$

Among them, w_{td} is the weight of feature *t* in text *d*; tf_{td} represents the times of feature *t* appears in text *d*; *N* represents the number of texts of training set; n_t represents the number of texts that appear feature *t* in all texts.

3.5. Support Vector Machines (SVM)

Support vector machine does not deal with the problem of high dimensional vector and sparse matrix data [16]. It is directly divided into two categories of text classification problem, effectively solve the problem of high dimension, and the accuracy rate is high, the generalization ability is good, but the selection of the kernel function and the parameters is more difficult than other methods, it usually rely on several experimental or empirical judgment.

Support vector machine was originally aimed at two classification problems. But in fact, it is usually faced with multi-class classification problems, such as short text classification. In solving the problem of multi classification, there are two ways to solve the problem: (1) decompose the multi classification problem into two classification problems; (2) design the multi-class classifier directly. At present, most researchers choose the first way in the text classification.

1) One-Versus-One (1-v-1)

One-Versus-One SVM method is to construct a SVM classification model from two randomly chosen classes of text, *k*(*k*-1)/2 SVM classification models are needed for the *k* class of the text. When classifying texts, the voting rule is used to select the category of the text when the category gets the highest score. The disadvantage of this method is that the lager *k* is, the more classifiers need to be constructed, and the longer training time will be cost. The computational efficiency is much lower than the One-Versus-Rest SVM method, so this method is not usually used.

2) One-Versus-Rest (1-v-r)

One-Versus-Rest SVM method is to construct a SVM classification model from two classes, one is divided from a class of text, and the other is divided from the rest classes of text, *k* SVM classification models are needed for the *k* class of the text. When classifying texts, it will be classified into the category corresponding to the maximum classification function value. The advantage of this method is that it is simple and easy to implement, and the disadvantage is the need to re-label all the samples, training time will increase with the increase of *k*, and two constructed classifications are very asymmetric, the negative samples are more than the positive samples.

3.6. Semi-Supervised Learning

Semi-supervised learning is a key problem in the field of pattern recognition and machine learning. It is a kind of learning method, which is a combination of supervised learning and unsupervised learning. It mainly considers how to use a small number of labeled samples and a large number of unlabeled samples to train and classify the text. There are five main algorithms: the algorithm based on probability, methods for modifying the existing supervision algorithm, methods that directly depend on clustering hypothesis, methods based on multi-view, and the methods based on graph [17].

But in fact, semi-supervised learning also has its limitation. Sample data without noise interference is used by most of the semi supervised learning method, but in real life, most data has the noise interference. It is often difficult to get the sample data without interference. If we use the sample data with noise interference, with iterative learning,

error samples will increase, the accuracy rate of classification will decline. So before semi-supervised learning, noise reduction is needed for sample data.

3.7. Improved Semi-Supervised Learning Algorithm

3.7.1. Algorithm Description

This paper used the SVM classification model to train the pretreated short text training set. There were many unlabeled samples in the database which used in this paper, so the semi-supervised learning algorithm was also used to iterate training set and calculate the similarity between the samples, and used the similarity between the samples to label the unlabeled samples until all the samples in the training set were labeled completely. Then input the testing set into the model to compute the effect of the classification model.

3.7.2. Algorithm Steps

Step 1: Input the training set T_r;

Step 2: Use the SVM classifier to train the labeled sample set P to gain the classification model M_1;

Step 3: Use M_1 to train the unlabeled sample set Q to label the samples;

Step 4: Iterative train unlabeled sample set Q until all samples were labeled;

Step 5: Reuse the completely labeled training set T_r' to gain the improved classification model M_2;

Step 6: Input the training set T_e into M_2;

Step 7: Output the result.

4. Experiment Result and Effect Analysis

4.1. Experimental Data

The data we use in this paper is from Sina weibo and Tencent weibo, which involves in politics, economy, education, entertainment and science & technology. There are 56985 short texts in the short text corpus we build. The distributions of the specific data are shown in the following table. Among them, we randomly choose 70% as the training set, 30% is used as the test set. We also need to choose 30000 short texts randomly as unlabeled texts to do the semi-supervised learning training.

Table 1. Experimental Data

	Training set	Testing set	Total
Politics	7973	3417	11390
Economy	7981	3421	11402
Education	7968	3415	11383
Entertainment	7990	3424	11414
S&T	7977	3419	11396

4.2. Evaluating Indicator

In this paper, the accuracy rate (Precision), recall rate (Recall) and F1 value are used to evaluate the classification results. Specific calculation formula is as follows:

$$Precision = \frac{\left|\cap(Prediction\ Set, ReferenceSet)\right|}{\left|Prediction\ Set\right|}$$

$$Recall = \frac{\left|\cap(Prediction\ Set, ReferenceSet)\right|}{\left|ReferenceSet\right|}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Among them, all predicted data are in the *PredictionSet*, and all the right data are in the *ReferenceSet*.

4.3. Experimental Results Contrast

In this paper, the KNN algorithm is compared with the algorithm which is proposed in this paper. The results are shown in the following two tables. Table 2 shows the precision, recall and F1 results of KNN algorithm, and Table 3 shows the results of the algorithm which is proposed in this paper.

Table 2. The Classification Result of KNN Algorithm

	Precision	Recall	F1
Politics	67.75%	69.08%	68.41%
Economy	62.37%	63.84%	63.10%
Education	61.96%	63.28%	62.61%
Entertainment	71.25%	72.89%	72.06%
S&T	65.56%	68.54%	67.02%

Table 3. The Classification Result of the Algorithm Proposed in this Paper

	Precision	Recall	F1
Politics	76.16%	77.67%	76.91%
Economy	72.79%	74.12%	73.45%
Education	68.73%	70.19%	69.45%
Entertainment	80.49%	81.77%	81.12%
S&T	73.98%	75.27%	74.62%

From the data, we can see that the effect is significantly improved when we use the algorithm, which is proposed in this paper, compared with KNN algorithm in politics, economy, education, entertainment and science & technology, every aspect has been improved.

Next, we use a line chart to show the effect of the classification, so that we can see the improvement more directly.

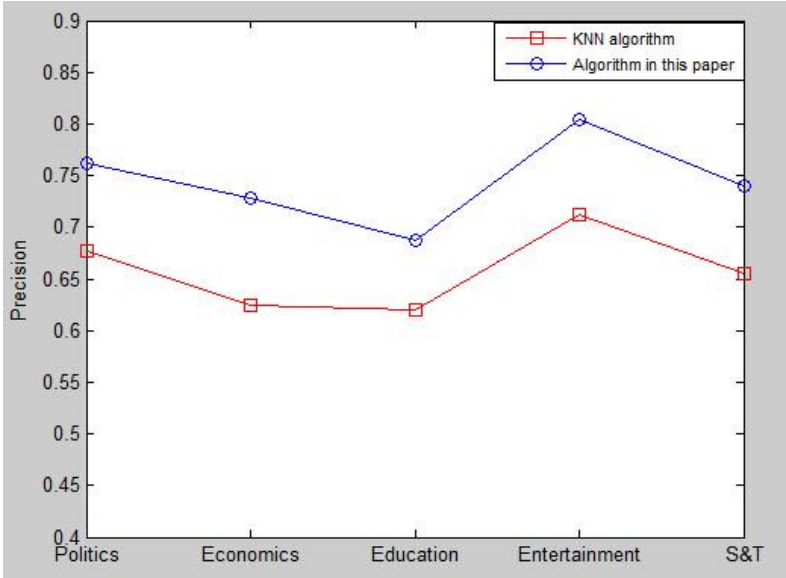


Figure 1. Precision Results Comparison

From Figure 1, we can see that the precision of the algorithm, which proposed in this paper, is better than the precision of KNN algorithm in politics, economy, education, entertainment and science & technology, the highest precision is 80.49%.

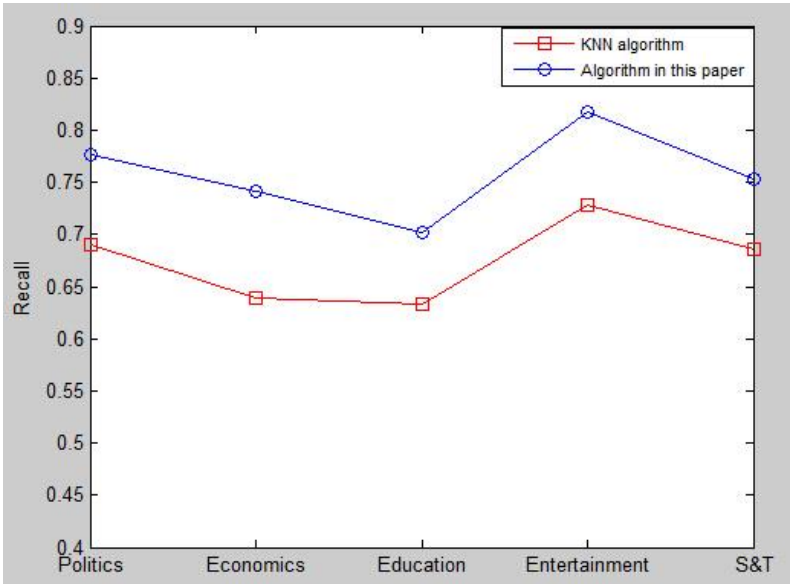


Figure 2. Recall Results Comparison

From Figure 2, we can see that the recall rate of the algorithm in this paper is also better than the recall rate of KNN algorithm in politics, economy, education, entertainment and science & technology, the highest recall rate is 81.77%.

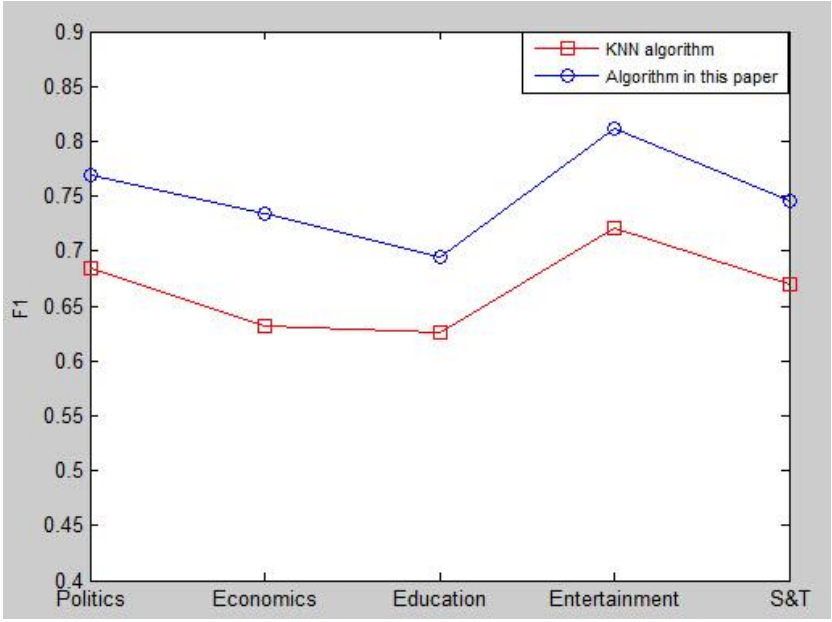


Figure 3. F1 Results Comparison

From Figure 3, we can see that the F1 value of the algorithm proposed in this paper is better than KNN algorithm in politics, economy, education, entertainment and science & technology. Using SVM and semi-supervised learning method, the classification effect is improved.

5. Conclusion

There are many unlabeled samples in the short text, less features and higher irregularity. The general classification model is not very good for short text classification. In this paper, we combined the SVM and semi-supervised learning method to learn and label the unlabeled samples in the short text, the effect of the trained classifier is better than the general model. The experimental results show that this method does improve the classification effect of short text.

However, the use of vector space model has a disadvantage that the extraction of the characteristics of the order and the relationship have been ignored, these elements (the order and the relationship between the characteristics) have a certain effect on the classification of short text. Nowadays, the amount of data in a short text has been great, and the efficiency is still a hot spot in the short text classification algorithm. We will do more research on the efficiency of the classification algorithm of the short text, and continue to improve the classification algorithm to get better results.

Acknowledgment

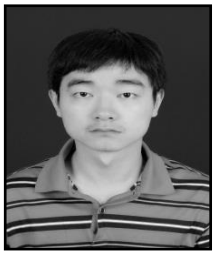
This paper is a revised and expanded version of a paper entitled “Short Text Classification Algorithm based on Machine Learning” presented at AITS 2015, Harbin, China, August 21-23, 2015. This work was funded by the National Natural Science Foundation of China (61373134, 61402234), and by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE) Korea. It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (No.KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). Professor Jin Wang is the corresponding author.

References

[1] D. Naiyang and T. Yingjie, "New method in data mining—support vector machine (SVM)", Science Press, vol. 16, no. 2, **(2004)**, pp. 113-126.
[2] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano and S. Li, "Incremental Support Vector Learning for Ordinal Regression", IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 7, **(2015)**.
[3] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman and S. Li, "Incremental learning for ν-Support Vector Regression", Neural Networks, vol. 67, no. 15, **(2015)**.
[4] T. M. Cover and P. E. Hart, "Nearest neighbor Pattern Classification", IEEE Transaction on Information Theory, vol. 13, no. 1, **(1967)**, pp. 21-27.
[5] F. Yunjie and L. Huailiang, "Research on Chinese short text classification based on Wikipedia", Modern library and information technology, vol. 24, no. 3, **(2012)**, pp. 47-52.
[6] L. Xiaojun, Z. Meng and B. Xiao, "Short text Classification Method Based on Concept Network", Computer Engineering, vol. 36, no. 21, **(2010)**, pp. 4-6.
[7] Z. Xiaohua, "Research on feature word weight algorithm in KNN text classification", Shanxi, Taiyuan University of Technology, vol. 37, no.13, **(2010)**, pp. 25-33.
[8] G. Salton, A. Wang and C. S. Yang, "A Vector Space Model for Information Retrieve", Journal of the America Society for Information Science, vol. 18, no. 11, **(1975)**, pp. 613-620.
[9] Z. Huayu, S. Zhengxing and S. Fuyan, "A Chinese text automatic classification system based on vector space model", Computer Engineering, vol. 27, no. 2, **(2001)**, pp. 15-21.
[10] Yin C., "Towards accurate node-based detection of P2P botnets", Scientific World Journal, vol. 42, no. 15, **(2014)**, pp. 146-158.
[11] C. Yin, M. Zou, D. Iko and J. Wang, "Botnet Detection Based on Correlation of Malicious Behaviors", International Journal of Hybrid Information Technology, vol. 6, no. 6, **(2014)**, pp. 96-104.
[12] Veeraswamy A., "A survey of feature selection algorithms in data mining", International journal of advanced research in technology, vol. 18, no. 1, **(2011)**, pp. 108-117
[13] Z. Kunhong, "Research on Chinese text automatic classification based on VSM model and feature selection algorithm", Jiangxi Normal University, vol. 54, no. 19, **(2011)**, pp. 254-262.
[14] M. L. Samb, F. Camara, S. Ndiaye, Y. Slimani and M. A. Esseghir, "A Novel RFE-SVM-based Feature Selection Approach for Classification", International Journal of Advanced Science and Technology, vol. 43, **(2012)**, pp. 27-36.
[15] A. Allahyar and H. S. Yazdi, "Data Ranking in Semi-Supervised Learning", International Journal of Advanced Science and Technology, vol. 53, **(2013)**, pp. 1-10.
[16] Y. Yang and X. Liu, "A re-Examination of Text Categorization Methods", Proceedings ACM SIGIR, Nanjing, China, October 23-29, **(1999)**, pp. 42-49.
[17] Banerjee S., Ramanathan K. and Gupta A., "Clustering Short Texts Using Wikipedia", In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, **(2007)**, pp. 787-788.

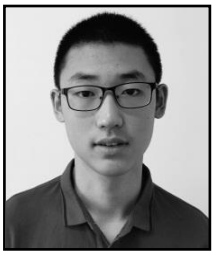
Authors

Chunyong Yin, is currently an associate Professor and Dean with the Nanjing University of Information Science & Technology, China. He received his Bachelor (SDUT, China, 1998), Master (GZU, China, 2005), PhD (GZU, 2008) and was Post-doctoral associate (University of New Brunswick, 2010).He has authored or coauthored more than twenty journal and conference papers. His current research interests include privacy preserving and network security.



Jun Xiang, received his bachelor degree in 2014 from Nanjing University of Information Science & Technology. He is studying for his master's degree in Nanjing University of Information Science & Technology. His main research interests include data mining, data security and privacy protection.

Hui Zhang, obtained his B.E. degree in the Measurement and Control from Nanjing University of Information Science and technology, China in 2014. Currently he is a graduate student at the School of Computer and Software of Nanjing University of Information Science & Technology. His current research interests are in big data mining and recommendation technology.



Zhichao Yin, is studying in Nanjing No.1 Middle School. His current research interests include network security and mathematical modeling.

Jin Wang received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the Computer and Software Institute, Nanjing University of Information Science and technology. His research interests mainly include routing method and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.