Analysis and Research of Several Problems of Bad Short Message Filtering System

W.F. Du
School of Mathematics
Physics and Information Engineering
Jiaxing University
Jiaxing, China

G.X. Chen
Hebi Civilian-Run and State-Subsidized Occupation Education
Management Center
Hebi China

***Abstract*—the spread of bad message seriously affects the social ethos and disrupt the normal life order of people. It has considerable practical value to research and develop the filtering technology of bad short message. Two problems in text classification are studied in this paper, which can be used in the bad short message filtering. The first is the application of clustering method to purify unreliable corpus. Experiment shows that the method is quite obvious on purification effect of unreliable data; the second is about a little improvement of word weight index IDF.**

Keywords-message filtration; unreliable corpus; vector space model; IDF

I. INTRODUCTION

With the rapid development of China's mobile communication business, because of its cheap price and convenience the short message service won the majority of users of all ages. SMS began to be called "the fifth media" after newspaper, broadcast, TV and the Internet. It is based statistics that China's mobile phone users have exceeded one billion for the first half of 2012. There is more than 8 hundred billions text messages are sent and 8 ten billion yuan of income every year. More than one billions messages had been sent in the user's thumb every day. It is more than 1/3 of the world total message. However, along with the "thumb economy" explosive growth at the same time, people are unbearable about the bad message. Almost all mobile phone users have suffered from similar SMS harassment.

With the proliferation of bad messages, it not only affects our normal communication order, but also poisons the social atmosphere. It disrupts the normal order of socialist market economy of our country and the influence is not conducive to the economic construction, social development and progress.

To control the transmission of harmful messages, both the authorities need to formulate relevant laws and regulations, also need certain technical means to take the identification and filtering on the bad message. These paper researches two problems in text classification can be used in bad short message filtering. The first is the purification of unreliable data sets; the second is about an improvement on TFIDF word right metrics.

II. THEORETICAL BASIS

A. Vector Space Model

Vector space model is proposed by Salton [1]etc. in the 60's of the last century, it is a branch of information retrieval methods in recent years. Vector model proposes a suitable framework for partial matching in recognizing the limitations of binary weight of the Boolean model. In the Boolean model, there are only two states of relevant and irrelevant for the text relative to the expression, but in the vector model, the text relative to the query string has a similarity value. Because the vector space model is built on the foundation of normal mathematics, the model has very wide use in the field of information retrieval. With the characteristics of items and their corresponding weights representing text information in vector space model, so the application premise is the central meaning of text can be reflected by the lexical information. In recognition, the degree of correlation between the text and the query request is described by the vector operation. If the new text and user interest model in text information filtering also use vector form, then the vector space model will also apply to the field of information filtering.

B. Clustering

Clustering is the process of object classification according to certain requirements, without any prior knowledge about the classification in this process, there is no category label object exists, and the similarity between objects as the only classification criteria, so it belongs to the category of unsupervised classification[2]. Clustering is to process a collection of objects grouped into consists of classes of similar objects. Clustering is an old problem, it is accompanied by the emergence and development of human society and the deepening of human to understand the world, we must distinguish between different things and understanding of similarities between things. From the perspective of human understanding of the world, clustering is an important means of human knowledge, people in the know strange things, while recognizing that a certain number of things after the individual, would not consciously use clustering method to classify the. And then summing for each category, properties are in each category for all individuals, were guided by the classification to the future to. Therefore, in the process of human understanding of the objective world, applying clustering thought earlier in the classification of thinking. Can say, the corresponding induction method to cluster thinking in logic,

deductive method and classification corresponds to the logic of thinking.

C. The Similarity Between Classes

Suppose two classes C, D having m, n samples respectively, i.e.

$$C = \{c_1, c_2, \cdots, c_m\}$$
$$D = \{d_1, d_2, \cdots, d_n\}$$

The similarity between classes could be inducted by formula (1) [3]:

$$\mathrm{sim}(C, D) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{sim}(c_i, d_j)$$

Among them, $\mathrm{sim}(c_i, d_j)$ is similarity between vector c_i and d_j, it can be represented by the cosine of the angle.

$$\mathrm{sim}(c_i, d_j) = \frac{c_i d_j}{|c_i||d_j|} \quad (2)$$

Partition of training set and test set

For supervised learning, the corpus will be divided into training set and testing set. The training set is for the learning of algorithm and the test set for assessing the effectiveness of the algorithm.

For partitioning the training set and test set, there are mainly two kinds of methods in recent days[4]: holdout method and k-fold cross validation method. The data will be divided randomly into two parts—the training data and the testing data in holdout method. Generally, 2/3 of data is as training data, and the other 1/3 as test data. Using training data derived classification model, its classification accuracy on test data is as the final classification accuracy. K-fold cross validation method will divide the data randomly into k roughly equal subsets of data S_1, S_2, …, S_k, training and testing are repeated k times. In the ith process, S_i as the test data, the other subsets are used as training data. The final classifier classification accuracy take k test average classification accuracy. This method is applied to a small amount of the original data, and then can not be directly applied to maintain method.

D. Close Test and Open Test

To classify the test set using the classifier, the classification results can make the evaluation on the performance of the classifier. The test has close test and open test. In the close test, evaluation set is part of the training set, or the training set itself; In the open test, the test set is independent and identically distributed with training set[5]. In general, the results in close testing have little significance, open test is mainly applied in classification.

E. TFIDF

It usually has a higher accuracy in weighted system of measuring words weight with a weight value replacing Boolean representation. In order to deal with professional words and general words which are both high frequency word, TFIDF (Term Frequency Inverse Document Frequency) is used in the field of text mining to measure the weight of a word, the formula is as following:

$$\mathrm{TFIDF} = \mathrm{TF} \times \mathrm{IDF} \left(\mathrm{IDF} = \log\left(\frac{N}{n}\right)\right) \quad (3)$$

Where TF denotes the entry's weight in the text, N said the total number of training set text, n to represent text contains the entry number. As for formula (3), there is the following intuitive explanation:

(1)The more the word appears in the text, the word is more representative of text, so the weight is greater;

(2)The word appears in the more text, the word's distinguishing ability of text classification is lower, so the weight is less.

III. PURIFICATION OF UNRELIABLE CORPUS

A. Delete of Similar Messages

The corpus collecting from the internet is from different site. There are a large number of text messages are the same or similar, these similar messages will not add information to the classifier, otherwise they will interfere with the learning of classifier, and increase the space and time complexity of the classifier. This paper presents a method to delete similar messages.

If we use the method of file comparison, only the exactly same message can be checked out. And through the analysis of the corpus, we have found that there is just slightly different in the wording of many short messages, but the contents are almost the same.

Therefore we vectorize each message using VSM, and measure the similarity between the given two vectors with the cosine of the angle given by formula (2). Setting a threshold, if the similarity exceeds the threshold value, delete one message corresponding to the vector. Through the experiment, we found that the threshold setting to 0.95 is more appropriate, most similar messages could be removed. After this step, a total of 330 similar messages deleted.

B. To Reduce The Effects of Error Classification Messages

In addition, we also found that, many of these SMS message was incorrectly classified. For example, some of the messages is obviously a normal SMS, but wrongly classified into yellow message class. However, it is a time-consuming work to read all the messages one by one to determine its category artificially.

How to purify impure corpus, this paper presents an approach. Assuming that the original corpus is divided into n classes C_1, C_2, …, C_n. Then using the method of clustering, the corpus is clustered into n classes. However, the classes produced by clustering method have no class labels, how to find the corresponding relationship between clustering generated class with the original classification? Assuming that the clustering method generating n classes D_1, D_2, …, D_n. We can apply the similarity between classes to determine which original class clustering D_i is corresponding to, applying formula (4):

$$J = \arg\max_j \mathrm{sim}(D_i, C_j) \quad (4)$$

Then the clustering D_i is corresponding to the original class C_J.

The results of the similarity between classes are shown in table 1:

TABLE I. THE SIMILARITY BETWEEN CLASSES.

	C_1	C_2	C_3
D_1	0.2182	0.0231	0.0178
D_2	0.0164	0.2520	0.0219
D_3	0.0145	0.0174	0.2684

Among them, C_1, C_2, C_3 represent normal SMS, yellow SMS and reactionary SMS respectively. From table 1 it can be drawn that clusterings D_1, D_2, D_3 are corresponding to the previous classes C_1, C_2, C_3 respectively. Therefore, we divide the corpora twice in fact:

$$\pi_1 = \{C_1, C_2, C_3\}$$
$$\pi_2 = \{D_1, D_2, D_3\}$$

The purification method of corpus is shown as following, if a message in the original classification is classified as some class, it is classified as the same class in clustering, we keep the message, or we abandon it.

How on earth is the effect of this kind of purification method? The close test and open test are carried out before and after the purification of the corpus. There are only less than two thousand messages in the purified corpus; the data is less, so the 10-fold cross validation method is used in the test.

Precision, recall and micro average under close test are shown in table 2:

TABLE II: COMPARISON OF THE ACCURACY RATE UNDER CLOSE TEST.

class	Pre-purification			Post-purification		
	precision	recall	micro average	precision	recall	micro average
normal SM	87.9%	90.3%	88.58%	97.6%	93.2%	95.46%
yellow SM	90.2%	81.3%		95.5%	94.4%	
reactionary SM	87.8%	95.5%		92.6%	99.5%	

From the result under close and open tests, the purification effect of unreliable data is rather good.

IV. IMPROVED IDF

The main idea of IDF is: if a text contains more entries, i.e. n is larger, then IDF is smaller, the ability the entry distinguishing category is lower. If the number of messages in a class C containing the entry is m, while the total number of messages in other classes containing the entry is k, apparently all the messages number containing the entry is $n = m + k$. When m is large, n is also higher, the value of IDF will be smaller according to formula (3), it means that the ability the entry distinguishing category is not strong. But in fact, if an entry frequently appears in the text of a class, it could be a good representative of the characteristics of text of this class, such an entry should be given a higher weight, and selected as characteristic word of this type of messages to distinguish with other messages, this is the shortage of IDF.

Aiming at the shortage of IDF, the literature [6] proposed some improvement. Suppose that the total number of messages is N, the number of messages containing the entry is n, the number of messages in a class C containing the entry is m, then there is formula(5) about the entry in class C:

$$\mathrm{IDF} = \log\left(\frac{N}{n} \times m\right) \quad (5)$$

If in addition to class C, the number of messages containing the entry is k, then the formula can be transformed into formula (6):

$$\mathrm{IDF} = \log\left(\frac{N}{m+k} \times m\right) \quad (6)$$

Literature [6] also proved the properties of the formula:

(1)IDF is a strictly increasing function of m;

(2)IDF is a strictly decreasing function of k.

The above properties actually expresses the following meanings, the number of messages in a class C containing the entry is large, and the number of messages in other classes containing the entry is small, the entry can represent the feature of class C, and has good ability of distinguishing categories.

But in some cases, some entries only appear in one class, i.e. $k = 0$, then

$$\mathrm{IDF} = \log\left(\frac{N}{m+k} \times m\right) = \log(N)$$

No matter how is the number m of messages in a class C containing the entry is, the value is always $\log(N)$, which is contrary to the fact. It should be that the number m is larger, the value is more important. Therefore this paper revises the formula in literature [6]:

$$\mathrm{IDF} = \log\left(\frac{N}{k+1} \times m\right) \quad (7)$$

The relationship between the value of IDF and m, k is as follows:

Set $f(m) = \frac{N}{k+1} \times m,\ m_1 > m_2$

$$f(m_1) - f(m_2) = \frac{N}{k+1} \times m_1 - \frac{N}{k+1} \times m_2 = \frac{N(m_1 - m_2)}{k+1}$$

Because $m_1 > m_2 > 0, k \geq 0, N > 0, f(m_1) - f(m_2) > 0$.

The improved IDF is still strictly increasing function of m.

Set $f(k) = \frac{N}{k+1} \times m,\ k_1 > k_2$

$$f(k_1) - f(k_2) = \frac{N}{k_1+1} \times m - \frac{N}{k_2+1} \times m = \frac{Nm(k_2 - k_1)}{(k_1+1)(k_2+1)}$$

Because $k_1 > k_2 \geq 0, m \geq 0, N > 0, f(k_1) - f(k_2) < 0$.

The improved IDF is still strictly decreasing function of k.

So the improved IDF can still maintain the two properties, then the entry can represent the feature of class C, and has good ability of distinguishing categories.

V. THE ANALYSIS AND ASSESSMENT OF THE RESULTS

The data is acquired based on the classification method of kNN before and after improvement of IDF, the results of precision and recall are shown in table 3:

TABLE III. COMPARISON OF THE ACCURACY RATE BEFORE AND AFTER IMPROVEMENT OF IDF.

class	Pre-purification			Post-purification		
	precision	recall	micro average	precision	recall	micro average
normal SM	87.5%	83.1%	79.18%	91.4%	84.6%	82.59%
yellow SM	83.2%	60.4%		83.7%	68.8%	
reactionary SM	65.9%	95.9%		68.2%	96.6%	

As can be seen from table 3, a slight improvement in the indexes of precision, recall and micro average is gotten after the IDF is improved. Through analysis there are only 8 feature words meeting the above conditions $k = 0$ under the limitation of 1000 features words. The proportion is very small, so the effect is not too obvious.

VI. CONCLUSIONS

In view of the great harm of bad message flooding, this paper discusses two technique problems of text classification which can be applied to bad message identification and filtering. Combined with the clustering analysis method, we realize the purification of unreliable corpus; the experimental results show that this method is quite effective. In addition, a reasonable improvement about IDF is put forward. If the proportion of the feature words appearing in only one class is higher, the improvement will have a certain effect.

REFERENCES

[1] G. Salton, A. Wong, C. Yang. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11)

[2] Gao Xinbo. Fuzzy Cluster Analysis and its Applications[M]. Xi'an: XiDian University Press, 2004

[3] Li Bichen, Shao Meizhen, Huang jie. Pattern Recognition and its Applications. Xi'an: XiDian University Press, 2008

[4] Han J, Kamber M. Data Mining: Concepts and techniques[M]. Beijing: Higher Education Press, 2001

[5] Li Jiabin. Text Classification Research Based on Cross Cover Algorithm [D]. Anhui University Master Degree Dissertation. 2007

[6] Zhang Yufang, Peng Shiming, Lv Jia. Improvement and Application of TFIDF Method Based on Text Classification[J]. Computer Engineering, 2006, 32(19)