# Application of Bayes Classification method in mobile phone spam short message filtering system

Lu Li[1*], Xiangui Xue[2]，Tao Li[3]

[1]School of computer and information, Qiannan Normal College for Nationalities,Duyun 558000
[2]Institute of mathematics and statistics, Qiannan Normal College for Nationalities, Duyun 558000
[3] School of computer and information, Qiannan Normal College for Nationalities,Duyun 558000

**Abstract:** The paper discussed the use of Bayes classification method in filtration system of short message spam(SMS). The method can classify the content of SMS, thus realizing effective filtering. Finally the paper carried out the result analysis and the appraisal of the Bayes classification model, which testified the model has some actually feasibility and extensibility.

## 1 Introduction

At present, a large number of spam messages have affected the normal life of people, so it is necessary to filter these spam messages. But in the filtration system, the traditional filtering methods are not thorough, which affects the users' life to a certain extent. In view of this situation, the filtration system used Bayes classification method to further improve the filtering system.

## 2 Material and Method

Bayes classification is based on the classification of Bias's theorem. The algorithm can predict the possibility of a sample belonging to a certain class of members. The samples are then assigned to the category with the highest probability. The traditional blacklist filtering system will filter directly when spam messages are sent to the blacklist of mobile phones. Otherwise, it is not. When the phone number is not on the blacklist, the Bayes classification model is used to identify the content of the message. After identification, if the message is not spam, the system will notify the user to read directly. Instead, the system displays the prompt information to the user ,but whether to read the message ,which is decided by the user. If the user chooses to read, the system will notify the user to read, otherwise the message will be filtered out.[1,2]

### 2.1. The realization process of Bayes classification program

(1) The filtration system reads the training samples and gets the statistics of all kinds of messages.

(2) The system reads the word dictionary to process the training sample text by word ,and which can get the corresponding DF value of every word. And then put the value into the corresponding database.

(3) According to the characteristics of vector selection method and the DF value from big to small, the system selects the first 50 all kinds of features words to form a feature vector.

(4) The system reads the test sample text to test and analysis on the Bayes classifier.

(5) The system reads an unknown message to identify the message using the Bayes classifier and to give the test results.

### 2.2. Segmentation procedure

When the classification model is established, the filtration system must get each category feature vector by word segmentation and get rid of the non-Chinese character. The process is as follows [2,5]:

(1) The filtration system puts training message into memory, and uses an integer variable C to record the ASCII code corresponding to each reading character. Now, the system reads the first character.

(2) The system must give the the scope of C value.If the value is in the 19800-41000 (Chinese character code range of Chinese character set), the system will add the character to the string variable named temp,otherwise, add a space(char) to temp variable.

(3) Then,the system reads the next character, and repeat the second step, until all the characters are read .

### 2.3. The matching of Chinese information

The matching of information means that people make a feature word list in advance (referring to each feature vector table). Then, the words in the thesaurus match the text message. If the match is successful, it is considered that the short message contains the feature word. Otherwise, this message does not contain this feature word. In the experiment, the feature word list is placed in the record storage of the record management system (RMS).

The system reads every word from the record and then matches the text message. If successful, it means that the message contains the word, otherwise, it does not contain the word.

## 3 Results

### 3.1. Database design

The record storage is a text that contains recordset, which is equivalent to the table in the database. Each record in the record store can have different lengths and can store different data. Each item in the record store is called Record. Each record has a unique identifier called recordID. This identifier recordID can be used to retrieve a record from the record store. The first recorded recordID is 1, the second is 2, and the next record recordID is more than the recordID of the previous record. In implementation, two adjacent records do not necessarily have a continuous recordID, especially when a record is deleted.

Programmers access records storage by recording storage names. In this experiment, there are three records storage, namely prizeTable (winning SMS), sexTable (yellow text message) and wishTable (Blessing SMS).

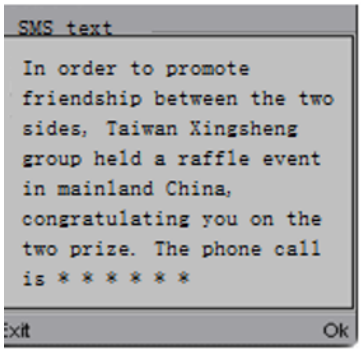


**Fig.1:** The input information simulation interface

### 3.2. The running interface of system

In J2ME, the input SMS is simulated through the interface, then the simple Bayes classification program is used to classify and identify short messages. The following is an analog interface. Users write short message content (as shown in Figure1), then press the OK key to call the classification recognition program, and the interface displays the returned identification information, as shown in Figure 2)[3].

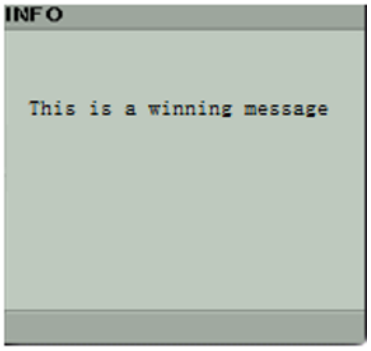


**Fig.2:** Identification information interface

## 4 Discussion

### 4.1. Methods assessment

The system uses the classification accuracy to measure. Classification accuracy is defined as:

$$\text{Accuracy}(M) = \sum_{t} P(t)\text{Accuracy}(M, t) = P(\hat{C}(t)) = C(t)$$

$$\text{Accuracy}(M, t) = \begin{cases} 1 & if \quad \hat{C}(t) = C(t) \\ 0 & else \end{cases} \quad (1)$$

Where,the C(t) is the actual class value of message t, $\hat{C}(t)$ is the calculation values classification model for T message,P (t) is the probability of message t (usually 1/n, n is the sample set size).[4,5]

### 4.2 Analysis of experimental data

The system collects only three kinds of messages to do the experiment. The system has good scalability. If we want to introduce different types of messages, the operation is very simple. Each message takes a certain number as the training samples, the other as a test sample. Figure3 is a schematic diagram of the classification results using the Bias classification model to test the message.

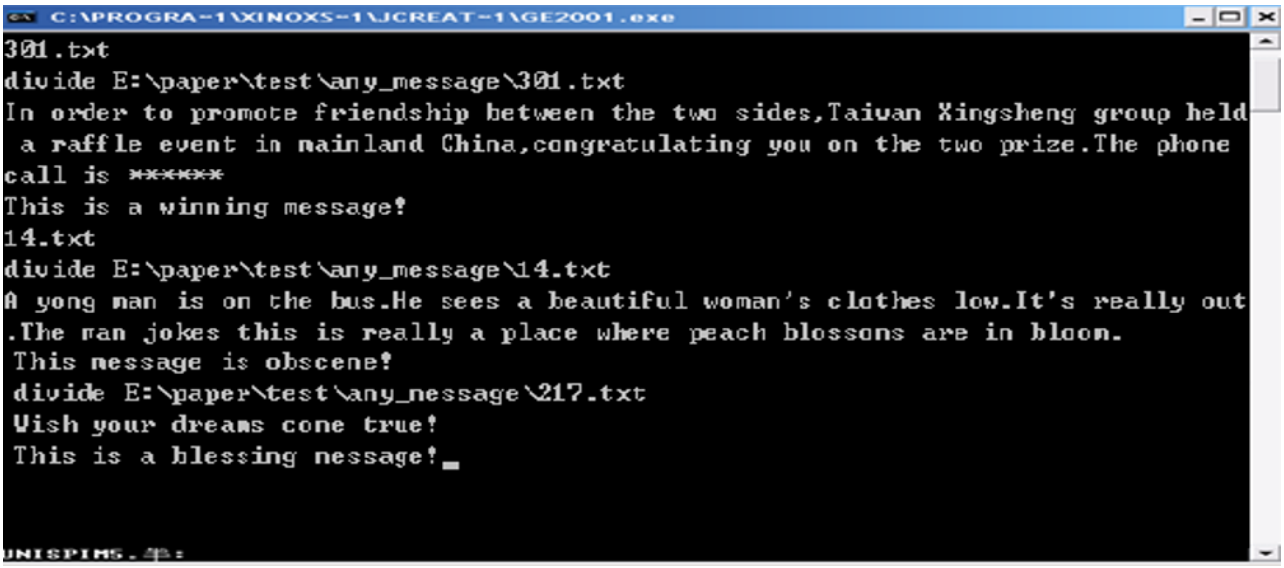


**Fig.3:** Schematic diagram of test results by using the Bias classification

## 5 Conclusions

This paper briefly discusses the application of Bayes classification in mobile phone spam message filtering system. The key process and algorithm implementation are given in this paper. And the results of the Bayes classification model are analyzed and evaluated. The result proves that this method is useful for filtering spam messages on mobile phones.

## Acknowledgements

The research was financially supported by Natural Science Foundation of Guizhou Provincial(Grant No.[2015]7723（Department of Guizhou LH words）) , Natural Science Foundation of Guizhou Provincial ( Grant No.[2014]7439（Department of Guizhou LH words）),and the research project of Qiannan Normal College for Nationalities (Grant No .qnsy201511 ).

## References

1. Zhang Y，Fu J-M.(2006) Identifying and tracebacking short message spam .The research and application of computer ,.245-247.
2. Cui X, Zhu S-F.(2006)Research on spam and anti spam Technology. Information and communication security.
3. Zhang Y-G,(2007) The research and implementation of mobile phone short message spam filtering system.Kunming:kunming University of Science and Technology master thesis.
4. Lei Y.(2011)Research on junk SMS multilevel classification technology based on Ensemble Learning.Chengdu:Sichuan University of Electronic Science and technology master thesis.
5. Huang Q, (2016)Research on correlation analysis method based on Bayes estimation and distribution . South China University of Technology.