Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STUDYING OF CLASSIFYING CHINESE SMS MESSAGES BASED ON BAYESIAN CLASSIFICATION

[1]LI FENG, [2]LI JIGANG
[1,2]Computer Science Department, DongHua University, Shanghai, China
E-mail:[1]Lifeng@dhu.edu.cn, [2]jznhljg@gmail.com

ABSTRACT

Although there are a lot of researches about e-mail spam filters, only a few focus on the issue for SMS (Short Message Service) system, especially in Chinese. In this paper, we proposed a two-layer filter model based on Naïve Bayes classifier utilizing both some traditional filter rules and content filter technical. The experimental results illustrate that the two-layer filter model can enhance the precision and efficiency of Bayesian classifier.

Keywords: *Naïve Bayesian; Bayesian classifier; Text classifier; SMS*

1. INTRODUCTION

As the number of mobile phone users continues to skyrocket, SMS is quickly becoming one of the fastest and most simple and economical forms of communication available. However, at the meantime of rapid development of SMS business, problems on information overload are also brought about. There are lots of SMS data, such as advertising, fraudulent, harassment messages, delivered to user opposing his or her need. Therefore, research on SMS classifying technology has important significance to assist people in dealing with SMS messages.

A variety of technical measures against spam email have been already proposed, such as decision trees, support vector machines (SVM), KNN, neural networks, etc. [1,2,7]. Most of them can effectively be transferred to the problem of mobile SMS filtering [4,6]. One of the most popular ones is the so-called Bayesian filter. Several independent researches have shown that content filtering based on the Bayesian classifier for short messages is surprisingly effective.

In this paper, on the basis of Bayesian filtering, we proposed a two-layer SMS filtering model. On the first level, we filter SMS through predefined rules, such as white and black listing, sending time, etc. On the second level, in addition to traditional Bayesian filtering using text of each message, we also consider several SMS specific features, for example text length, text structure. In this way, we can remarkably reduce the computation complexity and still keep the effective performance.

2. BAYESIAN CLASSIFIER

With solid mathematical theoretical basis and capability of comprehensive prior information and data sample information, Bayesian Classifier becomes one of research hotspots for current machine learning and text classifying.

1) Bayesian theorem

In probability theory, Bayes' theorem is to express

Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

how a subjective degree of belief should rationally change to account for evidence. The Bayes' theorem is used to calculate posterior probabilities, based on information that had been collected in the past. It represents a theoretical approach to inductive inference in statistical problem solving.

Mathematically, it was described as: for a test E, its sample space is S, B is an event of E, $A_1, A_2, \ldots, A_n$ is a partition of S, and $P(A_i) > 0$ $(i = 1, 2, \ldots, n)$, then

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} \quad , \quad i = 1, 2, \ldots, n. \tag{1}$$

2) Naïve Bayes classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with assumptions.

Suppose there is a set of m texts T = $\{T_1, T_2, \ldots, T_m\}$, where each text T_i is represented by a vector of n dimensions $\{x_1, x_2, \ldots x_n\}$. Values of x_i correspond to theA_1, A_2,..., A_n, respectively. In addition, there are k classes $C_1, C_2, \ldots, C_k$ and all texts belong to one of these classes. Given a sample of additional text t, it is possible to predict the class for t using the highest conditional$P(C_i|X)$, where I = 1, 2, …, k. The probabilities are derived from the Bayes' theorem:

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)}, \text{ i = 1, 2, …m.} \tag{2}$$

Conditional probability $P(X|C_i)$ can be figured out from

$$P(X|C_i) = P(x_1, x_2, \ldots x_n|C_i), \tag{3}$$

P (X) is constant for all classes. Only the product $P(C_i)\,P(X|C_i)$ should be maximized. $P(C_i)$ is the number of trained texts for class $(count\ of\ C_i)$ /(count of total). Given the complexity of the calculations $P(X|C_i)$, especially for large data sets, we conduct a naïve assumption of conditional independence between attributes. Using this assumption, it is possible to express $P(X|C_i)$ as a product:

$$P(X|C_i) = \prod_{j=1}^{n} P(x_j|C_i). \tag{4}$$

x_j represents the values for attributes in the text X. Probabilities $P(x_j|C_i)$ can be estimated from the trained data set.

Although the assumption of Naïve Bayesian approach is not accurate in reality, the fact is that it has achieved remarkably results. Many papers have shown that Naïve Bayesian filters can get a very good spam filtering results, showing this assumption to simplify doing little impact on performance [3][4].

3. SMS MESSAGE Filtering

Two layer filter model

Compared to general model with the Bayesian classifier, the model proposed in this paper adds another layer before content filtering. In this newly added layer, we firstly filter through the legit messages which accounting for the vast majority of the daily SMS messages by applying some rules, like black and white lists. Since this kind of filter tasks are easy to execute and are less time-consuming, we can save lots of unnecessary computation time. The new two-layer filter model is shown in Fig. 1.

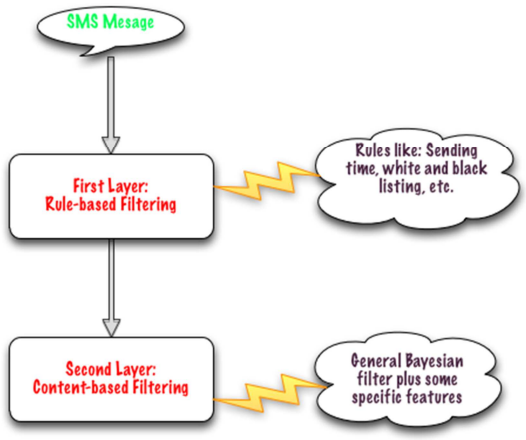


Figure 1 Two-Layer Bayesian Classifier Mode

A. Rule-based Filtering

Because different kind of SMS messages have specific features which sometimes are more

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

critical than its content. For example, people in your contact list seem much less likely sending fraud and phishing messages. Another example is that the time when people sending blessing messages to their friends are extremely close to the festival day.

We can use some application related rules to filter messages before use Bayesian filter. For some tasks, this scheme can reduce a lot of unnecessary computation. The most popular techniques used to reduce computation include the following ones.

a) Whitelist & Blacklist

The senders appearing in a black list are considered filter candidates, and their messages blocked. The messages from senders in a white list are considered legitimate, and thus delivered.

b) Sending time

When filter the blessing messages from normal texts, sending time is an extremely important factor to be considered firstly. While normal messages are sent every day, the blessing messages are only sent during the festival days.

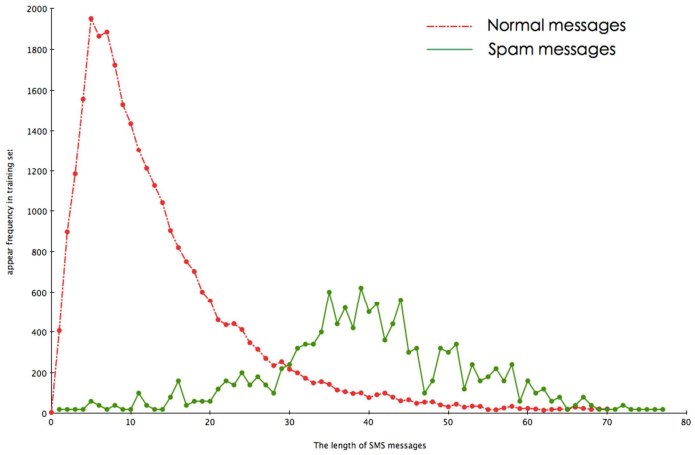


Figure 2 Length Comparisons Between Spam Messages And Normal Messages

c) Text length

In order to reduce the cost and to transfer as much information as possible, Spam messages are usually much longer than normal messages. The comparison of their length is shown in Fig. 2.

B. Additional SMS features

Traditional Bayesian filter just use 'bag-of-words' to represent a message. Some researches have been done to add weight on text words. But there are still some SMS related features we can utilize to enhance the precision. Following are some good candidates.

a) Specific punctuation

There are some punctuation marks whose probability of appearing in the normal characteristics of SMS messages and certain kind message is totally different. When filtering messages of these kinds, we can consider this term to be a feature as well.

b) Text structure

Some kind messages have special text structures comparing to others, such as symmetrical characteristic. By computing the length of different parts of a sentence separated by punctuations, we can test whether it is true or not.

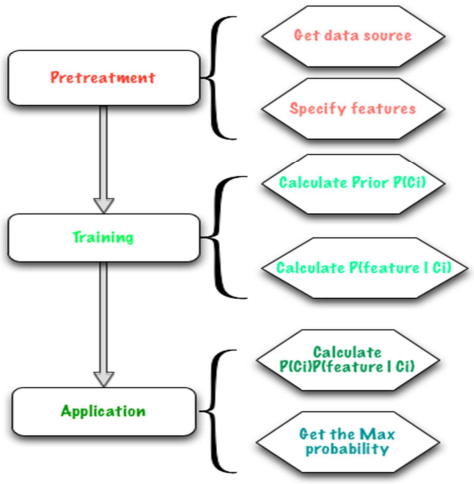
4. EXPERIMENT AND RESULT ANALYSIS

A. Experiment steps

To verify the correctness of the two-layer filtering model, we implement a SMS blessing messages filtering system on Android. Basically we consider filtering processing in three steps, as shown in Fig. 3

Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



.

Figure 3 Two-Layer Filtering System Process Flowcharts

a) Pretreatment Phase

Aim	**Do the necessary preprocessing work for Naïve Bayesian classifier**
Input	All the data to be classified
Output	Training Set and chosen features
Tasks	● Collect data ● Word segmentation ● Specify features and rules

▪ Collect data

The corpus in this experiment is collected in three ways:

1) Some SMS messages were extracted from the collection of the normal SMS messages and blessing SMS from user phones.

2) 2870 pieces of blessing SMS messages were collected from the Internet. We added the date property to each of them.

3) We also use the collection from NUS including 29533 pieces of Chinese SMS messages [12].

▪ Word segmentation

Because of the particularity of Chinese writing manner----there is no space between word and word; the first problem for Chinese text process is word segmentation [9,10]. Each word may include one or more Chinese characters. Some relevant statistical information shows that the average length of Chinese word is 1.59 [11]. Since Bi-gram model is a kind of N-gram model that is one of the most convenient language statistical models in natural language processing (NLP), we adopt Bi-gram as our word segmentation algorithm.

N-gram model is based on Markov theory model and can be described as follows:

We denote any text T as $T = w_1 w_2 \ldots w_n$. $w_i (1 \le i \le n)$ is a word and regarded as correlated between this word and preceding n words. The language statistical model is the probability of T.

$$P(T) = P(w_1 w_2 \ldots w_n)$$

$$= \prod_{i=1}^{n} P(w_i \mid w_{i-1} w_{i-2} \ldots w_{i-n+1}) \qquad (5)$$

In Bi-gram model, n=2, it is able to meet the needs of both computation precision and speed.

▪ Specify features and rules

After studying the characteristic of blessing SMS messages, we apply two rules and three main features here. Firstly we use sending time and messages' length as first filter layer features. Because there is no chance to sending a blessing message when it is not near the festival at all and there are dramatically difference in the length of normal messages and blessing messages. Secondly, besides the contents of SMS messages, we also consider the symmetrical structure and some particular punctuation like ";" and "!" as features of the messages.

b) Training Phase

Aim	**Generate the classifier**
Input	Features and training set.
Output	Classifier
Tasks	● Calculate the prior probability of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	each class ● Calculate the conditional probability of each feature in each class

In this step, we simply calculate the frequency of each class in the training set and the conditional probability of each feature, and then record them for later use.

c) Application Phase

Aim	**Use the classifier to filter data**
Input	Classifier and the data to be classified
Output	Classified result
Tasks	● Process the data to be classified through pretreatment phase ● Utilize the generated filter to classify them

Finally we feed the test data to the generated classifier and get the final map relation between test data and class.

B. Evaluation and analysis

We use the Naïve Bayesian classifier and the new two-layer filtering model to compare the experiments. The accuracy and efficiency of the algorithms were tested respectively. When comparing the efficiency, the time of Chinese word segmentation and read files were not included. The averaged data was taken after several experiments. The results are shown in Table1, Table2 and Table3.

Table 1 The accuracy of test results: traditional method

Training Samples	Precision	Recall	Fallout
500	94.78%	84.03%	5.03%
1000	95.24%	85.36%	4.36%
2000	97.03%	88.29%	2.87%

Table 2 The accuracy of test results: two-layer filtering method

Training Samples	Precision	Recall	Fallout
500	95.88%	84.03%	4.31%
1000	97.24%	85.36%	2.26%
2000	98.23%	88.29%	1.37%

Table 3 Efficiency Test Results

Training Samples	Traditional Method	The Improved Method
500	25s480ms	21s322ms
1000	29s763ms	23s128ms
2000	38s530ms	29s380ms

From Table1 and Table2, we can see that filtration rates had higher precision for the two-layer filtering method than the traditional Naïve Bayesian Classifier in the three different training scales. As Table3 shows, the time used for filtering the same blessing SMS messages by two-layer filtering method is less than the traditional methods. To sum up, the data comparison results show that the method introduced in this paper improves the overall performance of the Naïve Bayesian Classifier.

5. CONCLUSIONS

Based on Naïve Bayes classification techniques, this paper proposed a two-layer filter model. The experiment results show that the model can improve the classifier's precision and significantly reduce the misjudgment of legitimate messages. It is also superior to the traditional Naïve Bayesian filters on efficiency.

Our next work is to do further study on extracting

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

feature tokens automatically and optimizing the feature databases.

REFERENCES:

[1] Drucker, H, Vapnik, V., Wu, D. Support Vector Machines for spam Categorization. IEEE Transactions on Neural Networks, 10(5), pp. 1048—1054. 1999.

[2] Xiang,Y., Chowdhury, M., Ali, S. Filtering Mobile spam by Support Vector Machine. Proceedings of CSITeA-04, ISCA Press, December 27--29, 2004.

[3] Wu Jiansheng; Zhao Xingwen, "Improvement of Chinese spam filtering method based on Bayesian classification," Future Computer and Communication (ICFCC), 2010 2nd International Conference on, vol.1, no., pp.V1-765-V1-768, 21-24 May 2010.

[4] Belem, D.; Duarte-Figueiredo, F.; , "Content filtering for SMS systems based on Bayesian classifier and word grouping," Network Operations and Management Symposium (LANOMS), 2011 7th Latin American , vol., no., pp.1-7, 10-11 Oct. 2011.

[5] Gordon V. Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. Spam filtering for short messages. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007.

[6] Gordon V. Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. Feature engineering for mobile (SMS) spam filtering. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007.

[7] Q. Wang, yi Guan, X. Wang, "SVM Based Spam Filter with Active and Online Learning", In Procs of the TREC Conference. 2006.

[8] Christina V, KarPagavalli S, SUganya G. A Study on Email Spam Filtering Techniques. International Journal of Computer Applications (0975 - 8887), Volume 12 -, No.1. December 2010.

[9] Hui Jiao; Qian Liu; Hui-bo Jia, "Chinese Keyword Extraction Based on N-Gram and Word Co-occurrence," Computational Intelligence and Security Workshops, 2007. CISW 2007. International Conference on, vol., no., pp.152-155, 15-19 Dec. 2007.

[10] Nie Jian-yun, Gao Jiangfeng, and Zhang Jian etc., On the Use of Words and N-gram for Chinese Information Retrieval, Proceedings of the fifth international workshop on Information retrieval with Asian languages, 2000, pp. 141-148.

[11] Jian-Yun Nie, Jiangfeng Gao, Jian Zhang, and Ming Zhou. On the use of words and n-grams for Chinese information retrieval. In Proceedings of the fifth international workshop on on Information retrieval with Asian languages. 2000.

[12] Tao Chen and Min-Yen Kan. Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. CoRR abs/1112.2468. 2011.