

Chinese Spam Data Filter Model in Mobile Internet

Yitao Yang*, Runqiu Hu**, Guozi Sun**, and Chengyan Qiu**

* Department of Information Technology, Nanjing Forest Police College, Jiangsu, China

** School of Computer, Nanjing University of Posts and Telecommunications, Jiangsu, China

youngyt@gmail.com, B14040735@njupt.edu.cn, sun@njupt.edu.cn, B09040502@njupt.edu.cn

Abstract— With the rapid development of mobile Internet and the popularization of mobile intellectual terminals, learning, working and living were getting more and more efficient. While people were enjoying the efficiency if these modern products, large amount of spam information like advertising, pornography, gambling, fraud were flocking into people's daily life. The information mentioned above was mostly transmitted through networking data targeting any application. Different methods of detecting and filtering had been proposed by many researchers and scholars, among which the method of SVM based on content was the most popular. However, research focusing on network data had rarely been conducted. The paper proposed a networking data filter running on mobile terminals based on Bayesian Classification. It focused on the data of Chinese spam data detecting and filtering. In addition to the automatic learning of new rules, the filter also implemented a passive incremental learning. Experiments showed that the filter had a higher detection accuracy and an average level of system resource occupancy.

Keywords— Chinese Spam Data, Traffic Filter, Mobile Internet, Machine Learning, Bayesian Classification

I. INTRODUCTION

The rise of intelligent mobile terminals to people's lives to bring great convenience, so that work and study become more efficient. According to IDC's statistics[2], by the first quarter of 2017, intelligent mobile terminal equipment has taken up 96% of the mobile terminal device. In the report, it is not difficult to see that about 78% of the terminals are deployed with Android system and the quantity of App on Android system is much larger than that of other competitors. Early Android system is relatively fragile, liable to be attacked. Therefore, the protection of user's security and privacy are hard to be ensured. However, with the update of version, the security of the latest Android system is no less than iOS. However, the robust system does not provide many methods to solve the problem of harassment on users through spam data, because the data service (SMS) is provided by the communication operators[11]. The terminal is unable to prevent the sending of such messages. Although the operators have been making efforts to alleviate this problem, effect has not been seen obviously. In 2016, users received 20.6 spam messages per month on average (10.35 in 2008).

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution,

and reproduction in any medium, provided the original work is properly cited.

About 76% users have received various types of fraud information, causing a loss of 133 yuan per person, and an annual economic loss of about 91.5 billion Chinese Yuan.

Many researchers have proposed different types of spam detection and filtering methods, which can be divided roughly into two categories: content-based and non-content-based. Social network analysis is a typical non-content-based approach. Content-based methods include SVM, K-NN[4], logic regression algorithms[5], etc. In these methods, SVM is considered as the most suitable way of detecting spam data[7]. In recent years, machine learning algorithms have been increasingly introduced into the field[3]. The machine has become more intelligent and can learn new rules to filter the upcoming spam data. [9] proposed a spam message filtering model based on the Bayesian classification algorithm. The model can detect spam in short messages in the phone. Some researchers have also proposed a hybrid approach, such as content detection combined with challenge-response method[6, 10].

A report from a Chinese telecommunication operator pointed out that with the coming of the mobile Internet Age, the frequency of users using SMS service has met a sharp decline tendency. Compared with 2015, the use of SMS service of 52% of the users has decreased by 39%, while the use of intelligent terminal App data has increased by 60%. Spammers are also noticed this change, and the spam data was thus born. Spam data is the data which transmits information containing advertising, pornography, fraud and other content through the IP network. The destination of the data may be an App installed smart mobile terminals. It is difficult for the operator side to effectively detect and filter these spam data, so the best deployment location is intelligent terminals. At present, Apps like 360 Mobile Assistant is mainly functioned as the resistance of mobile phone viruses, Trojans, as well as filters based on the blacklist of single machine. The data detection and filtering has not been conducted. The paper designs a mobile data filter based on Android, which is based on the bottom of the networking services to filter all the App networking data, detecting and filtering the Chinese spam data using machine learning algorithm with feedback learning mechanism. The more the data has been filtered, the more knowledge it will obtain, and the higher the accuracy of detection will be reached. The remainder of this paper is organized as follow: Section 2 introduces Bayesian algorithm; a mobile data filtering model

will be presented in Section 3. The Section 4 shows a series of experiments and A conclusion is drawn in Section 5.

II. BAYESIAN ALGORITHM

Bayesian Classification[1] is abstracted from traditional probability theory and is also based on the theory. It particularly handles statistical problems with uncertainty. Bayesian classifier has a solid mathematical theory as well as ability to integrating prior and data sample information, making it one of the hottest researching fields in current machine learning and data mining. It's essential idea is to estimate the probability of given documents belonging to each class using the joint probability of words and categories definition and formula related to Bayesian algorithm.

Definition 2.1. A set of all possible outcomes of a random trial E is called the sample space of the random trial E (denoted as S). The element of the sample space, that is, each result of E , is called the sample point. A subset of the sample space of E (denoted as S) is a random event of E , and is referred to as an event.

Definition 2.2. Assume E is a random event, and S is its sample space. For each event A of E , a real number is given, denoted as $P(A)$, and is called the probability of occurrence of event A .

Definition 2.3. Assume A and B are two events, $P(A) > 0$, then

$$P(B|A) = P(BA)/P(A)$$

Theory 2.1. (Total Probability Formula) Assume the sample space of trial E is S , A is any event of E , $B_1, B_2, \dots, B_n$ is a partition of B and $P(B_i) > 0$ ($i = 1, 2, \dots, n$), then

$$P(A) = P(A|B_1) * P(B_1) + P(A|B_2) * P(B_2) + \dots + P(A|B_n) * P(B_n)$$

Theory 2.2. (Bayes' Theorem) Assume the sample space of trial E is S , A is any event of E , $B_1, B_2, \dots, B_n$ is a partition of B and $P(B_i) > 0$ ($i = 1, 2, \dots, n$), then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i)P(B_i)} (i = 1, 2, \dots, n)$$

III. SPAM DATA FILTERING MODEL BASED ON BAYESIAN ALGORITHM

A. Classification Algorithm

From the perspective of contents, spam data filtering can be considered as a binary classification problem, where the networking data is divided into two classes: spam and legitimate[1]. Bayesian classification algorithm is a widely-used classification algorithm. It has been applied to the text classification. By calculating the probabilities of a text belonging to each category, the text is classified into the class with the highest probability. The data can be converted into text form, and the text data needs to be pre-processed into other form of information that can be identified by computers. The text can be annotated with a set of words distributed independently, and the probability of the i th word in a given class C text is expressed as $P(w_i|C)$. Assuming that each word is randomly distributed in the text, that is, the word does not

depend on the length of the text, location of the other words in the text, as well as other textual content. For a given category C , the probability of the text D (made up with words w_i) showing up can be expressed as:

$$P(D|C) = \prod_i p(w_i|C)$$

The probability that document D belongs to class C can be written as:

$$P(C|D) = \frac{P(CD)}{P(D)}; P(D|C) = \frac{P(CD)}{P(C)}$$

Therefore, the following equation is conducted:

$$P(C|D) = \frac{P(C)}{P(D)} P(D|C) = \frac{P(C)}{P(D)} \prod_i P(w_i|C)$$

Combining with Theory 2.2, then:

$$P(C|D) = \frac{P(D|C)P(C)}{P(D|C)P(C) + p(D|C^c)P(C^c)} = \frac{P(D|C)P(C)}{\prod_i P(w_i|C)P(C) + \prod_i P(w_i|C^c)P(C^c)}$$

Now there are only two independent classes: S and $\neg S$ (spam data and normal data). Every element(data) here is either spam or not. With the equation above, $P(S|D)$ can be computed as following:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D|S)P(S) + P(D|\neg S)P(\neg S)} = \frac{P(D|S)}{P(D|S) + \frac{P(\neg S)}{P(S)}P(D|\neg S)}$$

let $k = \frac{P(\neg S)}{P(S)}$, then

$$P(S|D) = \frac{P(D|S)}{P(D|S) + kP(D|\neg S)} \quad (1)$$

B. Spam Data Filtering Model

In this Section, a series of definitions will be given. At last, an estimation model of spam data probability will be presented as the theoretical basis of the spam data filtration system.

Definition 3.1. Total data set $O = S \cup N$, where

S : {messages | known spam data, the set size $|S|$ is denoted as B };

N : {messages | known normal data, the set size $|N|$ is denoted as G }.

For spam data filtering, there are only two categories: spam data and normal data. The classification algorithm is carried out based on the data content, so the word segmentation is the first step before applying the classification algorithm. Chinese word segmentation is different from English word segmentation. The smallest unit is not a word but a single character.

However, the smallest unit with certain semantics is still word. In Chinese, there is no obvious segmentation between character and character or word and word, and this brings a lot of difficulties in Chinese word segmentation. [8] proposed a

method of word segmentation based on statistics. It calculates the credibility that a string is genuinely a word based on the frequency of neighboring words' co-occurrence. The method has ideal effects on Chinese, and is chosen to use for this passage. Each word (phrase) is written as token, and each data will get a token set after segmentation, which is defined as follows.

Definition 3.2. Segmentation function $p(x)$

$$p(x) = \{token | token \text{ are segmentation of } x, x \in O\}$$

Definition 3.3. Collection SPAM which is the segmentation result set from function $p(s)$: $\{token_s | token_s = p(s), s \in S\}$

Definition 3.4. Collection NORMAL which is the segmentation result set from function $p(n)$: $\{token_n | token_n = p(n), n \in N\}$. In order to calculate the frequency of the token in the SPAM and NORMAL sets, Definition 3.5 is given with two calculating functions as follows.

Definition 3.5. $bad(x)$ and $good(x)$ are two functions:

$bad(x)$ = the frequency of x in SPAM;

$good(x)$ = the frequency of x in NORMAL;

Definition 3.6. $\forall token (token \in SPAM \cup NORMAL)$, let $P_{spam|token}$ be the probability of any message containing the token is spam:

$$P_{spam|token} = \frac{r_b}{kr_g + r_b}, \text{ where } r_g = good(w)/G, r_b = bad(w)/B, k = B/G$$

Any set with this probability is denoted as P :

$$P = \{P_{spam|token_i} | token_i \in SPAM \cup NORMAL\}$$

Definition 3.7. For given data m , $T = \{token_i | token_i \in p(m)\}$, let $n = |T|$, T' is a sequential set:

$$T' = \{[P_{spam|token_i} - 0.5 | token_i \in T, 1 \leq i \leq n, P_{spam|token_i} \in P]\}$$

then the probability that m is spam is denoted as $P_{spam|m}$:

$$P_{spam|m} = \frac{\prod_{i=1}^n P_{(spam|token_i)}}{\prod_{i=1}^n P_{spam|token_i} + \prod_{i=1}^n (1 - P_{(spam|token_i)})} \quad (2)$$

C. Feedback Learning Mechanism

The object of filtering in the spam data filter is dynamic information flow, and feedback learning is an important component of the spam data filtration system. With time passing by, the form and content of spam data, and the individual needs of users are constantly changing, and the data filtering system needs to update the classifier according to the new changes and requirements.

To deal with the problem of feedback learning, two means are usually used: relearning and incremental learning. Relearning needs to update all intermediate statistics, to re-particularize all data and recalculate a prior probability, and the calculation work can be relatively heavy, making the problem very complicated. But for Bayesian incremental learning, only segmentation for new-joined data and calculation of statistical

word frequency is needed. This relatively reduced a lot consumption. Therefore, the system uses incremental learning to complete the system of feedback learning.

Denote data that the system will do incremental learning as Δm , $T' = \{token'_i = p(\Delta m)\}$ and do the following steps.

Step1. read one $token'_i$ from T ;

Step2. if $token'_i$ in collection SPAM then update $bad(token'_i)$;

Step3. if $token'_i$ in collection NORMAL then update $good(token'_i)$;

Step4. recalculate $P_{spam|token'_i}$ and update collection P ;

Step5. if exist $token'_i$ in T' has NOT been read then goto step1.

D. Filtering System Structure

The filter mentioned in Section 3.2 is deployed as a service on Android device. It can filter all traffic between terminal and Internet and detect the data into two classification as result: spam or normal.

The whole system can be divided into two phases. The first phase is initialization. This operation is performed only when the filter is run for the first time. It mainly finishes sample flow training, Chinese word segmentation, establishment of filter module based on the model described in Section 3.2, etc. Two types of with data will eventually be generated as two databases-normal data and spam data.

The second phase is a real-time operation. Whenever the Android system receives data flow, it will request the Bayesian classifier module. The classifier will calculate and return the classification based on the database obtained in the first stage: normal data or spam data. At this phase, there is another module-feedback learning module. When users actively trigger the module, it will perform the operation mentioned in Section 3.3 to do incremental learning on the users' input.

The entire system is deployed on Android devices, so the system reduces the functional module as much as possible to lower the system coupling degree. The code implementation is also optimized correspondingly on the mobile device. In Section 4, the analysis of the implementation effects and efficiency of the system is given.

IV. EXPERIMENTS

A. Filtering Model Parameter Test

Spam data samples Garbage flow samples comes from real data collected by China Mobile Communications Corporation, Jiangsu branch. The spam data set is selected in a month by the automatic filtering of the system. This experiment randomly selected 1000 data from the sample. Normal data samples Normal flow samples are collected from mobile phones of normal users. Up to 1000 sample data is randomly selected at the same time.

To achieve the spam flow model, three parameters needs to be determined, denoted as P_b , P_n and *threshold*.

Definition 4.1. See $P_{spam|token_i}$ in Definition 3.6, if $token^* \in SPAM \cap NORMAL$, then $P_{spam|token^*}$ is denoted as P_b .

Definition 4.2. If $token^{**} \notin SPAM \cup NORMAL$, then $P_{spam|token^{**}}$ is denoted as P_n .

Definition 4.3. Spam assessment criteria

if $P_{spam|m} > threshold$ then m is marked a spam message.

The parameters P_b , P_n and $threshold$ are determined by using the variable control method. P_b has a range of $[0.01, 0.99]$ with step length 0.01. P_n and $threshold$ have a range of $[0.1, 0.9]$ with step length 0.1. Therefore, there are a total of $99 \times 9 \times 9 = 8019$ groups of possible values to be substituted into the function. Then calculate the accuracy of the filtration by using 1000 test samples. After a comprehensive evaluation, the most appropriate combination of values will be selected. Because the number of possible combinations is too large, the results are grouped to enlarge the accuracy. For values of P_b , the results are divided into ten groups and the average accuracy rate of each group is then calculated. For P_n and $threshold$, the results are divided into nine groups. The average accuracy rate of each group is also calculated.

It can be seen from the three figures that with the increase of P_b and $threshold$, P_n decreases, and the accuracy of filtering gets higher and higher. Also, if a word never appears, it is highly possible to a normal word, for the words used in spam data are very common. In summary, the parameters (P_b , P_n , $threshold$) were selected as (0.99, 0.4, 0.9).

B. Filtering Effects Test

After training the sample data, the filter module starts its evaluation to the coming data flow. Assuming that the data to be tested has a total number of N , the result of a spam data filtration system is shown in Table 2. (Table 1. shows the meaning of identifiers used to calculate the evaluation parameters in Table 3.).

TABLE 1. Evaluation parameters

	Actual Spam	Actual Normal
Judged as spam	A	B
Judged as normal	C	D

*Note: the total number $N = A+B+C+D$

(1) Recall rate(Recall): $Recall = (A/(A + C)) * 100\%$. It stands for the rate of spam successfully detected. The higher the recall rate, the less the spam data "slipped through the net".

(2) Precision rate(Precision): $Precision = (A/(A + B)) * 100\%$. The rate of spam data correctly detected. It reflects the ability of the filter system to "correctly" detect the spam flow. The greater the precision rate, the less the number of non-spam data being misjudged as spam.

(3) Accuracy rate(Accuracy): $Accuracy = ((A + D)/N) * 100\%$ It is the rate of all data being correctly judged.

(4) Falling rate (Fallout): $Fallout = (B/(B + D)) * 100\%$ It is the rate of normal flow being misjudged as spam.

(5) Error rate (Error): . $Error = 1 - Accuracy$. It is, the rate of traffic to be tested.

(6) Wrongly reported rate (Missrate): $Missrate = 1 - Recall$. It is the rate of undetected spam data.

The distribution of test results is shown in Table 3. The performance evaluation of the system can be obtained from testing results distribution in Table 3. As is shown in the testing results, the performance of the system can fully meet the requirements of users. 91% of the high recall rate provides

the system with strong recognition ability. Although the fallout rate is up to 2% with a slight weakness, with the feedback mechanism of the system, the continuous learning will make it decrease gradually. In addition, the System time consumption is also fully within the user's affordable range, which can offer a real-time response.

TABLE 2. Testing results distribution (unit: piece)

	Actual Spam	Actual Normal
Judged as spam	455	10
Judged as normal	45	490

*Note: the total number $N = 1000$

TABLE 3. Filtering performance evaluation

Rec all	Precisi on	Accur acy	Fallo ut	Err or	Missr ate	System time consumpt ion
91.0 %	97.8%	94.5%	2.0%	5.5 %	9.0%	0.8ms/piece

C. Feedback Learning Test**Figure 1.** UI of Chat App

To test Feedback Learning Model(FLM), Chat, an Android APP developed for experiment, is used to transmit unencrypted data through mobile Internet. A message is considered as an unit of spam data in network traffic. Chat UI is roughly designed as Figure 1. In the figure, two messages with similar data content marked by red frame. These two messages has been detected as spam by filter so they are stayed in DUSTBIN. For testing, the first one is moved to INBOX manually.

**Figure 2.** The spam judge result after feedback learning

This operation will trigger the feedback learning mechanism. FLM update hash table mentioned in Section 3.3 and adjust the

probability of tokens segmented from the message. The tiny change will do effects on the spam assessment result. The Chinese messages which contain ADs keywords are sent to Chat after learning. In Figure 2., new spam messages are judged as normal message because of feedback learning. It testifies FLM works well.

V. CONCLUSIONS

This paper proposes an efficient Chinese spam data filtering system based on Android, which is based on the Bayesian classification algorithm, and is especially suitable for spam like fraudulent and advertisement spam. The spam data has a tendency to growing in fraud messages. Therefore, how to perform semantic analysis on the data content with the machine learning algorithm will be the focus of the future research.

ACKNOWLEDGMENT

This paper is supported by Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (No.PPZY2015A058), the Fundamental Research Funds for the Central Universities (No.LGZD201802) and National Natural Science Foundation of China (No.61772284).

REFERENCES

- [1] Bayes, T. 1991. *An essay towards solving a problem in the doctrine of chances*. 1763. MD computing: computers in medical practice.
- [2] China Mobile Phone Tracker, <http://www.itdcw.com/news/hangyeshuju/0504LO12017.html>: Accessed: 2017-08-23.
- [3] Deng, W.W. and Peng, H. 2006. Research on a Naive Bayesian Based Short Message Filtering System. (2006), 1233–1237.
- [4] Duan, L.Z., Li, A. and Huang, L.J.A. 2009. New Spam Short Message Classification. *Proc of the 1st International Workshop on Education Technology and Computer Science*. 168–171.
- [5] Li, H., Zhang, X., Yang, Y. and Sun, G. 2016. A Demonstration of Encrypted Logistics Information System. (2016), 525–528.
- [6] Mahmoud, T.M. and Mahfouz, A.M. 2012. SMS spam filtering technique based on artificial immune system. *IJCSI International Journal of Computer Science Issues*. 9, (2012).
- [7] Xiang, Y., Chowdhury, M. and Ali, S. 2004. Filtering mobile spam by support vector machine. *Proceedings of the Third International*

Conference on Computer Sciences, Software Engineering, Information Technology, E-business and Applications, pp. N. Debnath, ed. 1--4.

- [8] Xue, N.others 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*. 8, 1 (2003), 29–48.
- [9] Yang, Y., Hu, R., Qiu, C., Sun, G. and Li, H. 2017. A Spam Message Detection Model Based on Bayesian Classification. (2017), 424–435.
- [10] Zhang, J., Li, X.M., Xu, W.others 2011. Filtering algorithm of spam short messages based on artificial immune system. (Yichang, China, 2011), 195–198.
- [11] Zheng, X.X., Liu, C. and Zou, Y. 2010. Chinese Short Messages service spam filtering based on logistic regression. *Journal of Heilongjiang Institute of Technology*. 24(4), (2010), 36–39.



Yitao Yang He received the Bachelor Degree from Nanchang University in 2002 and the Master Degree from the School of Computing Science and Technology, Huazhong University of Science and Technology, Wuhan, in 2005. His current research interests focus on E-forensics, Security on Mobile Internet and IoT. He has authored or co-authored more than ten publications in which three are indexed by EI and one is indexed by SCI.



Master. Hu received the Bachelor Degree in School of Computer from Nanjing University of Posts and Telecommunications in 2018. His current research interests include Security in mobile communication and SDN.



Prof. Sun received the Ph.D. degrees from the School of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, in 2002. He is currently a professor of the School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, China. His recent research interests mainly include digital forensics, multimedia forensics, social network forensics, and digital investigation.



Master Qiu received the Master Degree in School of Computer from Nanjing University of Posts and Telecommunications in 2015. Her research interests in social media analysis.