Introduction of Fingerprint Vector based Bayesian Method for Spam Filtering

Bin Chen
Communication and Computer Network Lab, South China University of Technology
No.381, Wushan Road, Tianhe District, Guangzhou, Guangdong Province, CHINA
chenbin@scut.edu.cn

Shoubin Dong
Communication and Computer Network Lab, South China University of Technology
No.381, Wushan Road, Tianhe District, Guangzhou, Guangdong Province, CHINA
sbdong@scut.edu.cn

Weidong Fang
Communication and Computer Network Lab, South China University of Technology
No.381, Wushan Road, Tianhe District, Guangzhou, Guangdong Province, CHINA
fangwd@scut.edu.cn

ABSTRACT
With the development of the diversification of spam, it raises the difficulties and challenges to content-based spam filtering. To address this problem, this paper firstly introduced the statistical features of Email headers, and then proposed a method to use these features to improve Bayesian anti-spam filter. The selected Email-header features are presented as the fingerprint vectors, and then transformed to the input tokens to Bayesian filter. It shows that this method efficiently utilizes the messages embedded in Email headers and then improves the performance of the Bayesian anti-spam filtering.

1. INTRODUCTION
Over the past few years, many different approaches bad been released to against the spammers, e.g. SVM, Bayesian, rule-based and so on. Nevertheless, the spammers also know that technique, so they use many tricks to avoid being detected. Thus, the traditional Bayesian approach based on the content analysis cannot adapt this diversification well and result in the lower accuracy. Based on the statistic analysis about the behavior of spam corpus, we found that the Email headers of spam behave quite different from that of ham Emails; however such messages are often ignored by filterers. In this paper, we propose to use the fingerprint vector to represent the Email-header feature and the whole Email-body as token selections for Bayesian, to improve the traditional Bayesian filtering approach's performance and accuracy.

2. THE STATISTIC ANALYSIS ABOUT EMAIL HEADER FEATURE
Email header has abundant of attribute messages. Here lists some common fields in Email header: From, To, Cc, Subject, Replay-to, Date, Received and so on. From the observation to the real Email Corpus from our campus email server, we found that if the Email header has the following features, it has more suspicious to be a spam:

- **Wrong time zone.** For example, -600(EST), it is a false time zone, it should be EST (+800) for Chinese Emails.
- **Wrong time stamp**. An Email delivery may not be so long, if the time stamp between two MTA exceeds a limit, it has more probability to be forged (except for the transport mistake).
- **The mismatch of IP and the computer name.** If a user uses the MUA such as Outlook to send an Email, the Email would be firstly delivered to the MTA. In the delivery, the computer name may be unknown to the final MDA, but it is visible to the MTA
- **The mismatch of IP and Domains.** Usually, one MTA has only one domain, but the spammer also makes a trick to forge many false domains to escape the pursuit, so it usually is not correspond to the real IP.

Actually, according to the statistic [1] of our real email corpus, it shows that the mismatch percentage of IP and Domain in the spam is very high; it's about 86.6% on average while it is just 11% in the ham. We just take this for example to certificate that structured email header has much more important features to help us for spam detection.

3. FINGERPRINT VECTOR BASED BAYESIAN METHOD
The spammer is also changing their trick to evade this spam filtering. For example, they insert random characters to the email content to avoid the key word be detected, especially in Chinese spam. Because Chinese has no segmentation like English by the white space, so the traditional Bayesian can only take the whole sentence instead of the word as the token and its accuracy is decreased rapidly. Besides, more and more spam contains the media message like gif images, flash instead of the plain text, the traditional Bayesian and rule-based classifier can not treat this type of spam well. In order to address this problem, we propose an improved Bayesian method based on fingerprint vector to face with this challenge.

3.1 The Construction of the Fingerprint Vector
The fingerprint is a binary string used to sign an object. Usually, the fingerprint is made by a function set: $F = \{ f : \Omega \to \{0,1\}^k \}$, which Ω is available fingerprint object space, K is the length of the fingerprint.

CEAS 2007 – Fourth Conference on Email and Anti-Spam, August 2-3, 2007, Mountain View, California USA

The fingerprint vector is widely used in the detection of the similar documents, where the document *D* is regarded as a token sequence that can be represented by the fingerprints. For example, to the plain text, the token can be the characters, words and even the whole text lines; to the media, the token can be the pixels, size. The sequenced tokens in the document are called chunk or shingle. If the length of adjacent tokens is given by size w, it can be made a match between the document *D* and the chunk set D_w. Because the fingerprint is good to represent the message of whole document, we introduce the fingerprint vector to the spam filtering to cope with the diversity of spam.

The detail of the construction flow of fingerprint vector in our experiment is shown in fig.1.

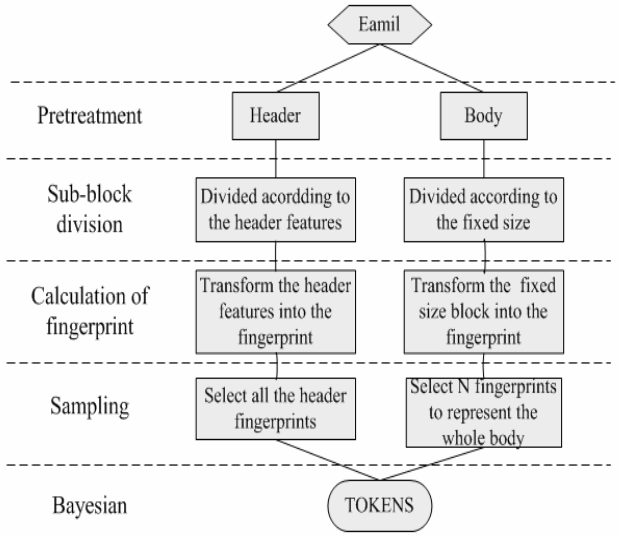

Figure 1. The flow chart of the construction of fingerprint vector

3.1.1 Pretreatment

As mentioned in section 2, Email header is a structured message, but Email body is unstructured, the content type maybe the plain text, html pages, multi-media and other MIME types. In order to make the classification more effective, an Email is divided into two parts: the header and the body, and is de-noised for each part. The purpose of this step is (1) to exclude the code that does not affect the text semantics, such as white space, HTML tag and so on; (2) to make the characters case insensitive; it means that all the characters are transformed into the unified pattern.

3.1.2 Sub-block Division

This step includes two different parts. Firstly, the high common spam features are selected from the Email header as followings: Subject, From, To, Received, date, Content-type, Content-transfer-encoding, Message-ID, Reply-to, and Charset. We use the LEX and YACC [2] technique to construct a vocabulary analyzer, for example:

UINT8: ([01]?[0-9]?[0-9]?|2([0-4][0-9])|5[0-5])

IPADDR: {UINT8}\.{UINT8}\.{UINT8}\.{UINT8}}

Then we used the vocabulary analyzer to parse the whole Email header, and obtain the representative Email header blocks. It is noticed that the length of sub-block will affect the accuracy of detection based the fingerprint vector. The sub-block is bigger, the accuracy of similar detection would be higher; but the sub-block will be more sensitive to the character noise. In our experiment, we find that when the length of the Email body sub-block is 6 to 8 bytes, the spam detection achieves the optimal result.

3.1.3 The Calculation of the Fingerprint Vector

The good fingerprint algorithm must meet the following two points: the fingerprint conflict rate and the calculation complexity should not be low. Rabin algorithm [3] is used to transform the sub-blocks into the fingerprints; the main idea is as followings.

Let $A = (a_1, a_2, ..., a_m)$ be a binary representation of a character sting, then the association between the string A and a polynomial $A(t)$ of degree $m-1$ with coefficients in Z_2 field is:

$$A(t) = a_1 t^{m-1} + a_2 t^{m-2} + ... + a_m \quad (1)$$

Let $P(t)$ be an irreducible polynomial of degree k, over Z_2 field. Here, the polynomial is:

$$P(t) = b_1 t^k + b_2 t^{k-1} + ... + b_{k+1} \quad (2)$$

With fixed P, the fingerprint of A is defined to be the polynomial:

$$f(A) = A(t) \bmod P(t) \quad (3)$$

Rabin fingerprint algorithm had been improved that the fingerprint conflict rate is very low, nearly 1% in 1 million Email body [4].

3.1.4 The Fingerprint Vector Sampling

The above statistics presented in the second section show that the spam feature is quite distinguished from some Email header fields, so the fingerprints from the Email header has greater weight, and all of them will be picked up in the fingerprint vector. For the Email body, we only choose the minimal n fingerprints to represent the whole body. Refer to the method that used to judge document similarity in [5], the minimal value is set to n = 80. Finally, the selected fingerprints are constructed as the tokens for Bayesian filtering method.

3.2 The Bayesian Filtering Method

Bayesian algorithm has been widely implemented for spam filtering and has made a great success on anti-spam. The main steps of the Bayesian algorithm are as followings.

1. Extract the independent string as a token from training corpus and calculate the word frequency, e.g. The number of this token respectively occurred in ham and spam corpus.

2. Use the word frequency to establish two hash table; one is hash table of ham, the other is hash table of spam, and store the mapping relationship of the token to the word frequency respectively.

3. Calculate the probability of each token in each hash table P :

P = the word frequency of the token / the length of the corresponding hash table.

4. Calculate the probability of a new Email that is spam when some tokens appear in it based the hash table of ham and hash table of spam. Let *A* stands for the situation that the Email is spam, $t_1, t_2, ..., t_n$ are tokens, $P(A \mid t_i)$ stands for the probability of a new Email that is spam when it has the t_i token. If $P_1(t_i)$ is the value of t_i in hash table of spam, $P_2(t_i)$ is the value of t_i in hash table of ham, then:

$$P(A \mid t_i) = P1(t_i) / [P1(t_i) + P2(t_i)] \quad (4)$$

5. Establish a new hash table named hashtable_Pro to store the mapping relationship of the token t_i to the probability $P(A \mid t_i)$.

When a new Email comes to the MUA, N tokens are taken from this Email and the value of each token is obtained by querying the hashtable_Pro. Let the individual token be $t_1, t_2, ..., t_n$ and the corresponding value be $P_1, P_2, ..., P_n$ in the hashtable_Pro, then calculate the probability of this new Email that is spam by the composite probability formula:

$$P(A \mid t_1, t_2, t_3, ..., t_n) = (P_1 * P_2 * ... * P_n) / [P_1 * P_2 * ... * P_n + (1 - P_1)(1 - P_2)...(1 - P_n)] \quad (5)$$

If the value of $P(A \mid t_1, t_2, t_3, ..., t_n)$ is greater than the predetermined threshold, this new Email will be classified as spam, otherwise it will be classified as ham.

And the filter which joined in the CEAS live-challenge competition, the threshold we preset is 0.52.

In our experiment, we also do the Bayesian optimization and the result smoothing using Robinson edge detection method that is implemented in Bogofilter [6].

4. CONCLUSION

Currently, the anti-Spam methods such as the traditional rule-based, content based Bayesian and SVM can not cope well with the diversity of spam. In this paper, we proposed an improved Bayesian based on the Email header feature and fingerprint vector. This method does not need word segmentation and decoding. According to the fingerprint algorithm, it is also convenient to transform the multi-media spam message that is expressed by binary stream, into the fingerprint vector. Thus, this method would be also useful in spam detection of binary formats.

The experimental results show that, it is possible to improve the performance of Bayesian filtering method that used Email header feature and fingerprint vector [1].

5. ACKNOWLEDGMENTS

This project is supported by the National 973 Program of China (No. 2003CB314805) and the National 863 Program of China (No. 2006AA01Z196)

6. REFERENCES

[1] Bin Chen, Shoubin Dong and Weidong Fang, Journal of Computational Information Systems 4:3(2008) 1205-1212.

[2] Levine Jobn R., Mason Tony, Brown Doug. Lex & Yacc.2003

[3] Rabin M.O. Fingerprint by random polynomials[R]. Cambridge: Center for Research in Computing Technology, Harvard University,1981

[4] LIU Jie, CHENG Xueqi. Fast Spam Detection Method under High-speed Network. Computer Engineering, vol.4,2006

[5] FANG Wei Dong, Network Spam Detection and Filtering Technologies. Ph.D. Thesis, South China University of Technology, China, 2007

[6] Bogofilter[DB/OL].Bogofilter,http://bogofilter.sourceforge.net/ , Feburary 1, 2007