Spam Detection Utilizing Statistical-Based Bayesian Classification

Xianghui Zhao[1], Yangping Zhang[2] and Junkai Yi[2,*]
[1]China Information technology security evaluation center, Beijing 100085, P.R. of China
[2]College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, P.R. of China
[*]Corresponding author

***Abstract*—Spam is one of the major problem of today's life because it causes a lot of extra expense both in network infrastructure and our individual life. Among those approaches developed to detect spam, the content-based detection technique, especially statistical-based Bayesian algorithm is important and popular. However, the basic Bayesian algorithm permits on assumption and estimation. In this paper, we proposed an improved method to increase the accuracy of the algorithm. Firstly, use actual priori probability instead of constant probability of spam. Secondly, expand the selective range and rule of tokens. Finally, add URLs and images into detection content.**

Keywords-spam; statistical-based bayesian classification; content detection

I. INTRODUCTION

Electronic mail (E-mail) is one of the most important and powerful communication methods in our business and personal lives. However, with the growth of email usage, the diffusion of spam is become deeper and deeper. Spam causes e-mail server engines to overload in bandwidth and server storage capacity [1]. Spam severely interfere with people's life, even give rise to financial loss and information security risks. Therefore, it's imminent to find an effective spam classification solution.

Many techniques against spam have been proposed (see below): keyword filtering, black-list, white-list, hashing, rule-based filter, statistical filter [2]. Among them, statistical filter (especially Bayesian filter) play a key role in anti-spam product. Spam recognition rate of an outstanding Bayesian recognizer can exceed 99.9%.

Paul Graham proposed Bayesian spam filtering algorithm in 2002. This algorithm get a lot of attention because of its simple, effective and intelligence. But the algorithm premised on two assumptions, uses a number of estimates in calculation process, thus affect the accuracy of the algorithm. In this paper, we present a spam detection method utilizing statistical-based Bayesian classification algorithm to address these limitations.

The remainder of this paper is organized as follows. Section 2 introduces the basic theory. Section 3 details our improved algorithm. Section 4 describes the spam detection method based on improved Bayesian algorithm. Section 5 executes experiments and presents the results. Finally, Section 6 draws the conclusions.

II. BASIC THEORY

A. Bayesian Classification Algorithm

The core of Bayesian filtering algorithm: If some words frequently appear in spam, but rarely appear in legal email, the message contain these words are more likely to be a spam [3]. In accordance to the occurrence of word in the mail, calculate the probability of a message is spam by Bayes theorem. The algorithm based on two assumption [3]: 1) The words present in an email are independent events. 2) There is no a priori reason for any incoming message to be spam rather than legal mail, considers p(s) = p(l) = 0.5.

B. Implement of Bayesian Classification Algorithm

The basic steps of Bayesian classification algorithm:

1) Gather a large number of junk mail (spam) and legitimate mail (legal), create spam and legal corpus.

2) Divide each mail into token set, create *hash-spam* table and *hash-legal* table, mapping each token to number of occurrence in each corpus.

3) Create *hash-probability* table, mapping each token t_i to probability of spam contains token t_i, which is calculated by (1):

$$P(A \mid t_i) = \frac{P_s(t_i)}{P_s(t_i) + 2P_l(t_i)}. \tag{1}$$

In (1), P (A | t_i) denotes the spam probability of an email which contains token t_i. P_s (t_i) indicates the probability of token t_i appears in spam corpus, which can be estimated by dividing the number of spam that contain this token by the total number of spam. $P_l(t_i)$ indicates the probability of token t_i appears in legal mail corpus, which can be estimated by dividing the number of legal mail that contain this token by the total number of legal mail. Here are some instrument about (1) we need to concentrate on: a) Doubling $P_l(t_i)$ to reduce misjudgment; b) Only count the token occurs more than 5 times to control the length of table. c) $P(A \mid t_i) \in [0.01,0.99]$ in order to avoid excessive influence on specific token.

4) Calculate the probability of any new mail is spam by hash-probability table.

a) Divide the new arrived mail into token sequence t_1, $t_2 \ldots t_n$ in accordance with step 2. Get P (A | t_1), P(A | t_2)…P(A | t_n) (Abbreviated as $P_1, P_2, \ldots P_n$) by hash-probability table. Set $P_i = 0.4$ while token t_i never occurs in hash-probability table. b) Select the top fifteen representative tokens, where representative is decided by $|P_i - 0.5|$ (Distance between P_i and 0.5), are used to calculate the probability that email is spam. c) Calculate email's spam probability using recombination probability formula:

$$P(A \mid t_1, t_2, \ldots t_n) = \frac{\prod_{i=1}^{n} P_i}{\prod_{i=1}^{n} P_i + \prod_{i=1}^{n} 1 - P_i}.$$

5) Any given email is determined as spam if $P(A \mid t_1, t_2, \ldots t_n)$ exceeds the threshold, for example, 0.9. End of the algorithm.

III. IMPROVEMENT OF BAYESIAN CLASSIFICATION ALGORITHM

A. *Improvement of Priori Probability*

Assume that P(S) indicates the priori probability of a new received E-mail is spam, then (1-P(S)) means the priori probability of legitimate mail. According to Bayesian formula, token t_i contribute to email's spam probability is calculated by (2)

$$P(A \mid t_i) = \frac{P_s(t_i)P(S)}{P_s(t_i)P(S) + 2P_l(t_i)(1 - P(S))}. \quad (2)$$

The second assumption of PG's Bayesian algorithm consider that the E-mail's spam probability is 0.5, in other word, P(S) = 0.5. This assumption permits simplifying the (2) to (1). However, the second assumption doesn't hold while the proportion of legitimate messages and spam is extremely high or low, therefor, the actual priori probability of spam must be consider, which means calculate the spam probability by (2).

B. *Improvement of Token*

Basically, Bayesian classification algorithm is based on probability statistic of tokens. It is an important way to improve efficiency of Bayesian by defining more effective selection rules of tokens.

*1) Introduce Phrase into Token:*To evade detection, spammers replace the small spam probability phrase with high spam probability word. Many experiments have shown that spam could evade detection more easily by replacement. It follows that the first assumption of PG's Bayesian algorithm doesn't stand in actual. Under this circumstance, we need introduce phrase into token.

Moreover, introduce the interrelation and expression of token itself. The E-mail content feature is shown in Figure 1. Vertex represent the word or phrase extract from E-mail; Numerical value ni represent the content expression force of the vertex; The weight on the edge wij represent relativity between two words or phrases.

2) Delete Rarely used Token: With accumulation of data, access efficiency of database goes down. We need to delete rarely used token to maintenance the database. Add "Count" and "LastUseTime" attribute in the three table to record token's usage information, which is used to clean-up the database.

3) Adjust occurrence of tokens: To avoid omissions and misjudgment, adjustment of tokens' occurrence is chosen. Increase occurrence of token in hash-legal table if legitimate mail is misjudgment which means inadequate attention of tokens. Similarly, in hash-spam table if spam is omitted. And update hash-probability table according to modified hash-legal table and hash-spam table.

4) Extract token from Chinese Email: It's very easy to extract tokens from English email, because sentence is composed of words and separated by space. Nevertheless, it's very difficult to extract token from Chinese whose sentence is composed of character, one or more continuous characters form a meaningful word which doesn't have specific delimiter. Segmentation algorithm is a good method to solve this problem. At present, there are three most used segmentation methods [4, 5]: string match based segmentation, statistic based segmentation and comprehension based segmentation.

Forward maximum matching algorithm (FMM) [5] and reverse maximum matching algorithm (RMM) [6] are common used string match based algorithm.

FMM: Divide string ABC into AB/C if $A \in W$, $AB \in W$, $ABC \notin W$, W is the dictionary.

RMM: Divide string ABC into A/BC if $C \in W$, $BC \in W$, $ABC \notin W$, W is the dictionary.

According to statistic based segmentation the following equation to calculate adjacent information M(X, Y) of adjoining character pair (X,Y).

$$M(X, Y) = \log \frac{P(X, Y)}{P(X)P(Y)}. \quad (3)$$

In (3), P(X, Y) indicates co-occurrence probability of (X,Y) in corpus; P(X) indicates probability of X in corpus; P(Y) indicates probability of Y in corpus. If M(X, Y) exceed threshold, consider "XY" as a word.

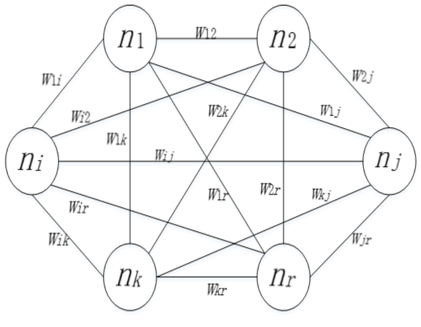

FIGURE I. EMAIL CONTENT FEATURE

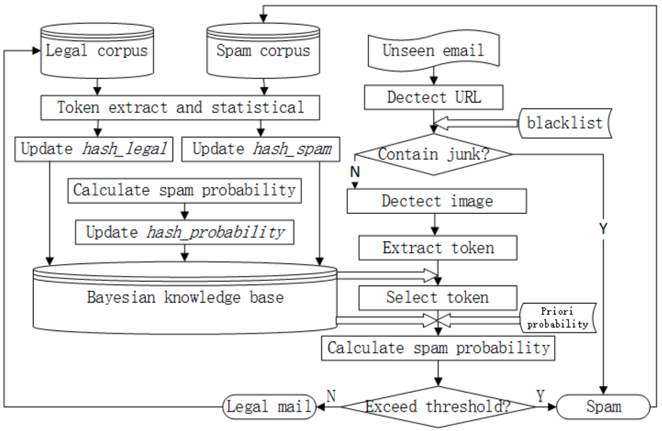

FIGURE II. BASIC STRUCTURE OF SPAM RECOGNIZER

C. Expand Detection Content

1) *URL Detection:* Spam often place a URL in the text to lure users to click. Blacklist [6] technology could be used to solve this problem. The email is judged as spam if the URL involve in blacklist. However, some cunning spammer would change IP address to evade IP-based blacklist filtering. It's very effective to confront this situation by behavior-based blacklist filtering.

Behavior-based filtering [7] check the source of mail according to sending behavior, include cluster and classification stage.

Cluster

Input tensor $M_{n \times d \times t}$ (n: The total number of IP address; d: The total number of domain; t: The total number of time slot.). M (i, j, k) indicates the number of email send to domain j in time slot k, by IP address i.

a) Calculate the $n \times d$ matrix as follows:

$$M^{'}(i, j) = \sum_{k=1}^{t} M(i, j, k).$$

b) Use spectral clustering algorithm [8] get IP address collection C: C$_1$, C$_2$,…,C$_n$, which meets the following Equation:

$$\mathrm{Y}_{i=1}^{k} C_i = n, \ C_i \ \mathrm{I} \ C_j = \Phi(i \neq j).$$

c) Calculate average value of each collection by (4), $M_c^{'}(i)$ is sub-matrix of $M^{'}(i, j)$.

$$C_{avg} = \frac{\sum_{i=1}^{|C|} M_c^{'}(i)}{|C|}. \tag{4}$$

Classification

Input $r_{1 \times d}$ vector of IP.

a) Calculate the similarity of vector r and collection C as follows:

$$sim(r, C) = \frac{r \bullet C_{ave}}{|C_{ave}|}.$$

b) Calculate similarity score of IP:

$$S(r) = \max_{c}(sim(r, C))$$

The IP address is spam source if S(r) exceed threshold.

2) Images detection: Image spam [9] is an obfuscating method in which the text of the message is stored as a GIF, JPEG, BMP or PNG image and displayed in the email. This prevents text-based spam filters from detecting and blocking spam messages. Currently, OCR scanning, source identification and fingerprint algorithm are used to solve the problem. The basic method of anti-spam product is extract script and color to judge the picture whether contains junk information or not [10].

IV. Spam Recognizer Based on Bayesian Classification Algorithm

The structure of spam detection system based on improved Bayesian is shown as Figure 2. First, create spam and legal corpus. Then, according to Bayesian training algorithm, divide each email into tokens, statistic the occurrence of each token. After that create hash-spam and hash-legal table. Meanwhile, calculate spam probability of each token contribute to, create hash-probability table. Use these three table create Bayesian knowledge base. When new email arrives, detect URLs, images and text, make the judgment with the help of Bayesian knowledge base and classification algorithm. At last, Feedback classified email into corpuses to prepare for new training process. After several feedback, start Bayesian training again if corpuses reach a certain fluctuation range, update hash-spam, hash-legal and hash-probability table to maintain timeliness of Bayesian knowledge base.

V. Experiment Result

We established the spam detection system based on improved algorithm, select 4000 email to test the system, including 2000 legal mail and 2000 spam. These emails are divided into 4 groups, select three of the group as training sample to classify the email, the others as test sample. , Cross validation by recombining the training and test sample several times, select the average value as the experiment result.

The recall ratio and accuracy is evaluation criteria of the experiment. Recall ratio reflect the ability of the system find spam, the higher of the recall ratio, the less of the evaded spam. Accuracy reflect recognizer's correct discrimination rate for all mail. In spam corpus, suppose that the number of email be judged as spam is A, the number of email be judged as legal is C. And in legal corpus the number of email be judged as spam is B, the number of email be judged as legal is D. The recall ratio P$_r$ and accuracy P$_a$ are calculated as follows:

$$P_r = \frac{A}{A + C} \times 100\,\% \qquad P_a = \frac{A + D}{A + B + C + D} \times 100\,\%$$

TABLE I. TABLE I. COMPARISON OF EXPERIMENTAL RESULTS

Recognizer type	**Evaluation index**	
	Recall ration[%]	***Accuracy[%]***
Our recognizer	97.8	99.2
PG's Bayesian	93.5	94.6

The experiment result is shown as Table 1, spam recognizer based on improved algorithm could increase the overall performance of filter effectively in contrast with PG's algorithm in recall ratio and accuracy.

VI. CONCLUSION

In this paper, we first research the principle and implement of Bayesian classification algorithm. Then, aim at the defects existing in the algorithm, make the improvement in priori probability, tokens' selection rules and detection content of spam. Finally, a spam recognizer based on improved Bayesian classification algorithm is designed. As the experiment results shows, the recalling ratio and accuracy of recognizer is improved significantly.

ACKNOWLEDGMENT

This research was supported by project (U1536116) fund by National Science Foundation of China (NSFC).

REFERENCES

[1] A. K. Sharma, R. Yadav, Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique, Conference On Communication Systems and Network Technologies (CSNT), (2015)1089 – 1093.

[2] N. Jatana, K. Sharma, Bayesian spam classification: Time efficient radix encoded fragmented database approach, Conference on Computing for Sustainable Global Development (INDIACom), (2014)939 – 942.

[3] Information on http://www.paulgraham.com/spam.html

[4] Huang Chang-Ning, Zhao Hai, Which is essential for Chinese word segmentation: Character versus word, 20th Pacific Asia Conference on Language, Information and Computation (PACLIC), (2006) 1-12.

[5] Chu Dong-Xue, Study on Chinese word segmentation algorithm, Conference on Manufacturing Technology and Electronics Applications (ICMTEA), (2014) 1536-1539

[6] West Andrew G., Aviv Adam J., Chang, Jian, Lee, Insup, Spam mitigation using spatio-temporal reputations from blacklist history, 26th Annual Computer Security Applications Conference(ACSAC), (2010) 161-170

[7] Ramachandran A., Feamster N., Vempala S., Filtering spam with behavioral blacklisting, 14th ACM Conference on Computer and Communications Security, CCS'07, (2007) 342-351.

[8] Bach, Francis R., Jordan, Michael I., Learning spectral clustering, 17th Annual Conference on Neural Information Processing Systems, NIPS 2003.

[9] S. Dhanaraj V. Karthikeyani, A study on e-mail image spam filtering techniques, Conference on Pattern Recognition, Informatics and Mobile Engineering, (2013) 49 – 55.

[10] Xiao Mang Li, Ung Mo Kim, A hierarchical framework for content-based image spam filtering, Conference on Information Science and Digital Content Technology (ICIDT), (2012)149 – 155.