Bayesian Filter Based on Anti-Spam Grid

LIANG YE, WEIMING ZHONG AND ZHIYONG XIONG
Department of Computer Engineering
Suzhou Vocational University
Suzhou, China
e-mail: {yel, zwm, xzy}@jssvc.edu.cn

PENG LIU
Grid Research Center, Institute of Command Automation
PLA University of Science and Technology
Nanjin, China
e-mail: milgrid@163.com

***Abstract*—Due to the gradual increasing of spam's harm to the society, the anti-spam technology has been the focus of attention of all sectors. In this paper, we bring forward the Bayesian filter technology based on Anti-Spam Grid, which can realize the cooperation among various servers and comprehensive utilization of grid resources, and ultimately improve the accuracy and efficiency of spam filtering.We analyze the Anti-Spam Grid, design and implement the grid-based Bayesian filter, and experimentally verify the effect of the filtering system.**

Keywords- Bayesian, Anti-spam, Grid

I. INTRODUCTION

E-mail now has been one of the main means of communication in modern society. And at the same time, spam is also growing rapidly[1].

Due to the gradual increasing of spam's harm to the society, the anti-spam technology has been the focus of attention of all sectors and relevant research has developed rapidly[2]. Because each anti-spam server works independently and the information cannot be shared, on the other hand, the spam attacks the entire Internet and the spam has many varieties, as a result, a single service is unable to filter the spam in an effective and timely manner[3][4].

Therefore, the way for the efficient and safe cooperation among various servers and resource sharing has been the focus of the study.

II. ANTI-SPAM GRID

A. System Architecture

In order to achieve cooperation among various servers and sharing resource on network, we designed a service-grid based anti-spam system. We have planned the anti-spam system architecture, known as Anti-Spam Grid. The use case diagram is shown in Fig. 1.

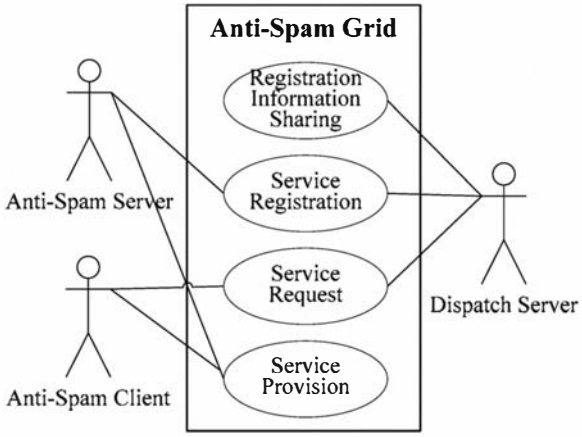

Figure 1. The Top Layer Use Case Diagram of the Anti-Spam Grid

B. Work Sequence of System

The main work steps for the Anti-Spam Grid are as follows:

a) The anti-spam server publishes the services of the server to a dispatcher, realizing the information sharing between dispatchers;

b) Once the anti-spam client connects with the Internet, the application to the dispatcher is made;

c) The dispatcher selects a server for serving the dispatcher in accordance with the load balancing or the selection of other strategies;

d) The client reports the signatures and Bayesian learning outcomes of e-mail to the server;

e) The server feedbacks the statistical information of the e-mail with signatures and the recently updated Bayesian learning outcomes of other clients.

The sequence diagram of the system is shown in Fig. 2.

978-1-4244-5517-1/10/$26.00 ©2010 IEEE

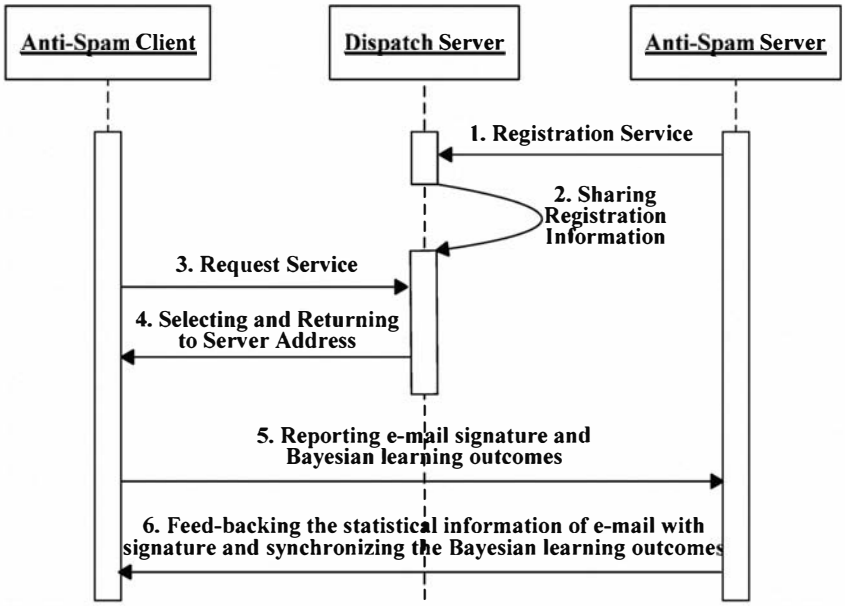

Figure 2. The Sequence Diagram of Anti-Spam Grid

III. DESIGN OF BAYESIAN FILTER

A. *Pre-processing Mechanism*

Before implementing the Bayesian filtering on spam, the pre-processing[5] must be made on the mail. When the mail is classified by the content-based approach, the mail is generally expressed as a multi-dimensional vector, and then a statistical method[7] can be used to calculate the possibility of the mail as spam. In the process when the e-mail is expressed as the multi-dimensional vector, the statistics on a lot of spam and normal mails must be made and the corresponding feature items shall be extracted according to specific criteria. After that, the feature item selected can be used as the dimension to express the mail as a multi-dimensional vector. Since the e-mail is semi-structured document, composed of e-mail header, message body and attachment, and the formats of mail vary. At the same time, the spam, to prevent the identification of the spam filter, may be intended to insert a number of useless characters and markings in the mail. Therefore, before expressing the mail, it needs to extract the mail body and implement the word segmentation, etc. The pre-processing on mails is also an important process in the spam filtering, the effective pre-processing can remove interference of spam and greatly enhance the filtering effects.

B. *Mail Expression and Word Segmentation*

In expressing the mail, generally, it needs to extract the mail body[6] from the mail first. In order to accurately extract the mail body, it needs to analyze e-mail header, which is very important, because much useful information can be obtained from the e-mail header. The content-type field in the e-mail header specifies the format of the file as the mail body. Where the format is plain text, the mail body remains unchanged; where the format is Html, all html tag will be removed from the mail and the plain text may be obtained. The adoption of the processing way can greatly reduce the interference of spam on the filter and improve the identification accuracy of the filter. Where the format is multipart, the boundary filed defines the boundary codes of different formats. Charset field describes the character set used in the mail body, the filter then applies the appropriate word segmentation technology based on the Charset field. Content-transfer-encoding field describes the way of mail code, the filter then transfer into the actual characters in accordance with the content-transfer-encoding field.

After the mail body is extracted, the plain text can be obtained, after that, it needs to implement the word segmentation on the mail in accordance with the character set defined by the charset field. The objects of the research in the paper are e-mails in English and Chinese, so only the English word and Chinese word segmentation are discussed here.

English word segmentation: For the mail in English, the space between words is deemed as the natural dividing line breaks and the punctuation is deemed as the semantics delimiter, that is to say, the English word segmentation is relatively simple. For obtaining the word list of a mail, it only needs to scan the English mail body without punctuation marks from the beginning to the end, in which the blank space between two spaces is identified as a word.

Chinese word segmentation: The segmentation technology is the basis for Chinese information processing; now there are many Chinese word segmentation methods available, such as maximum match based approach, optimal path method, feature database method, adjacency constraint method, artificial neural network method and dictionary-free method, etc. In this paper, ICTCLAS, an open-source Chinese word segmentation system of

Chinese Academy of Sciences is used for Chinese word segmentation.

C. *Bayesian Algorithm*

In the anti-spam field, Bayesian is a popular content-based filter algorithm. Naive Bayesian Algorithm, as a simple and effective classification algorithm, is applied to a lot of statistics-based machine self-learning document classification system[7]. In the document classification, for a document type C = (c_1, c_2, ..., c_n), document d = (w_1, w_2, ..., w_n), where c_1, c_2, ..., c_n is the document type, w_1, w_2 , ..., w_n is the value representing the features of document. In accordance with the Naïve Bayesian algorithm, the probability of a document d belonging to c_i is:

$$P(c_i \mid d) = \frac{P(c_i)\prod_{w_j \in d} P(w_j \mid c_i)}{P(d)} \tag{1}$$

Where, $P(c_i \mid d)$ is the probability of the document d belonging to c_i; $P(c_i)$ is the probability of any a document belonging to c_i; $P(w_j \mid c_i)$ is the probability of the word in c_i as w_j; and P(d) is the probability of any a document as d.

According to Bayesian hypothesis, it is assumed that the words and phrases in the text are independent of each other in determining the role of text type, the location of words in the document does not affect the type that the document belongs to. The formula for computing the probability of a word w_j belonging to c_i is as follows:

$$P(w_j \mid c_i) = \frac{n_j + 1}{n + |Vocabulary|} \tag{2}$$

Where, n_j is the number of w_j in c_i; n is the number of words belonging to c_i in the training corpora; |Vocabulary| is the number of w_j in the documents requiring classification.

For a specific document d, if the document belong to c_i, then $P(c_i \mid d)$ is the maximum. However, because P (d) is a specific value, the document d belonging to $C_{N.B.}$ is as follows:

$$C_{N.B.} = \max\ (P(c_i \mid d)) \tag{3}$$

IV. Implementation of Bayesian Filter

In the distributed Bayesian filter, when the mail server receives an e-mail, the fingerprint of the e-mail will be generated first, and then the query in the fingerprint database will be made, if the frequency of the occurrence of the fingerprint is greater than the threshold set, the E-mail can be marked as Spam. The server of the e-mail is also provided with the content-based filter, but the server can decide whether to open the filtering feature in accordance with efficiency or performance. After the end of the filtering and recording of the server, the mail then can be sent to the client.

The client is provided with the perfect content-based filter for filtering the mails without marks of the server. In filtering, in case any spam found, the fingerprints and the feature information used can be recorded into the log of the E-mail, and then the log is feed-backed to the server regularly or after being triggered. In case the client encounters gray e-mail, that is, the results for the mail by Bayesian filter and integrated scoring are inconsistent, the fingerprint information needs to be returned to the server. After that, the server sends the mail information requiring the identification to the network, and the accurate judgment on the mail can be obtained from the network, eventually, the judging results must be feed-backed to the client.

At the same time, the client must be responsible for reporting the user feedback information, in particular the mails wrongly judged. We present the following example for illustrating the specific processing of the system.

a) The mail server receives an e-mail

b) The mail server generates a fingerprint and implements the initial filtering on the mail, in case the mail is a spam, the relevant mark is given to the mail header and the log and fingerprints are recorded.

c) The mail is sent to the client.

d) The client then filters and marks the mails and implements the appropriate processing in accordance with the judgment results, in case the mail is a spam, the log information is recorded.

e) The log information of the spam is reported.

f) In case any suspicious e-mail is sent to the server, the request for judgment is made.

g) The server then checks the information of the suspicious e-mail on the Internet.

h) The network returns the query results.

i) The server gives the comprehensive assessment, record on the results and then sends back the relevant information to the client.

j) The user reports the misjudge mail.

k) The server records the misjudge information.

The sequence diagram on the system of mail processing is shown as Fig. 3.

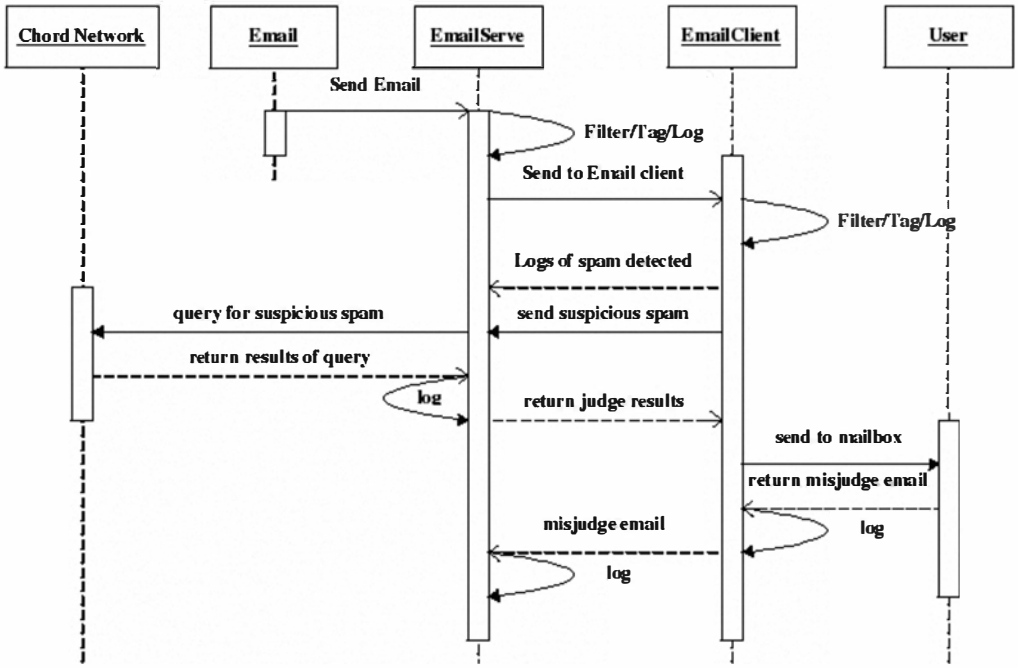

Figure 3. The Sequence Diagram of Mail Processing System

V. TEST ON SYSTEM EFFECT

We then mixed the real mails and simulated mails and used the mail sending server to send the normal mails and spam in proportionate, the experiment lasted nearly 200 minutes. And the test information on the normal mail and spam is shown in Fig. 4, from which, we can see that 95% of spam is found out and the misjudgment rate of normal mails is only 0.5%, the test shows that the effect of the filter is ideal.

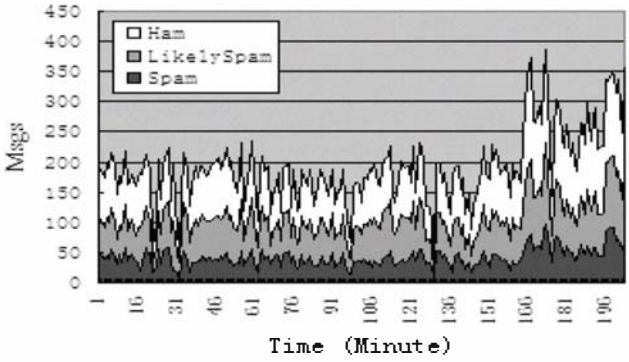

Figure 4. Statistics on Mail Filtering Test

VI. PROSPECT

The main feature and innovation of the system is the use of the distributed bayesian filter for the comprehensive utilization of grid resources.The system still can be further optimized, including the data synchronization of bayesian algorithm learning outcomes and others.At the same time, with the constant updating of anti-spam technologies and the rapid development of grid technology, the use of the grid for anti-spam will be perfect.

ACKNOWLEDGMENT

This work was supported by the China National Science Foundation with Grant No.60403043.

REFERENCES

[1] Spamassassin, http://au2.spamassassin.org/doc.html

[2] Plice, Robert K., Melville, Nigel P.: Toward an information-compatible anti-spam strategy, Communications of the ACM, v 52, n 5, p 128-130, May 1, 2009

[3] Kanaris Ioannis, Kanaris Konstantinos, Houvardas Ioannis, Stamatatos Efstathios : Anti-spam based on universal network measurement platform, International Journal on Artificial Intelligence Tools, v 16, n 6, p 1047-1067, December 2007

[4] Lu Xinjie, Chai Qiaolin, Ma, Li:Research and implementation of distributed spam detection System based on multi-agent system, Jisuanji Gongcheng/Computer Engineering, v 31, n 18, p 124-126, Sep 20 2008

[5] Jorgensen Zach, Inge Meador: A multiple instance learning strategy for combating good word attacks on spam filters, Journal of Machine Learning Research, v 9, p 1115-1146, June 2008

[6] Zhang L.E., Zhu Jingbo, Yao Tianshun: An evaluation of statistical spam filtering techniques, ACM Transactions on Asian Language Information Processing, v 3, n 4, p 243-269, December 2004

[7] Fan Yan, Zheng Cheng, Wang Qing-Yi, Cai Qing-Sheng, Liu, Jie: Using Naive Bayes to coordinate the classification of web pages, Ruan Jian Xue Bao/Journal of Software, v 12, n 9, p 1386-1392, September 2001