Natural Language Processing Technologies for Multi-Level Intelligent Spam Mail-Filter

Haiyan Kang and Xiaojiao Yuan

***Abstract*—To overcome the lack of existing mail filtering system, we designed a content-based message filtering system of multi-level intelligence. Using natural language processing technology, it denotes the E-mail content including attachments. First, it pre-processes the content of E-mail, including segmentation, feature extraction. Second, combining knowledge-base and expansion of the feature, it can form the vector. Corresponding categories vector in the database, two vectors similar degree of calculation determines the credibility of the message. Based on the above theory, with the Java EE 6+SQL Server 2005 platform, a mail filtering system is achieved. It can maximize the elimination of spam. The major features are following: 1) black /white list filtering. It can intercept white list blacklist e-mail messages released. 2) reverse DNS testing. it can effectively eliminate the anonymous e-mail attacks. 3) content-based message filtering. An accurate analysis of mail content can filter out suspicious messages. 4) fingerprint recognition. It can mimic the biological concept of fingerprint identification to complete the identification of spam. 5) user-personalized filtering. The user independently designed filter program. 6) intent detection. It can detect the content URL connection in email. Experiment shows mail filter system can play a very good effect on spam filters.**

***Index Terms*—Spam mail-filter, Chinese word segmentation, mail classification, privacy protection, information security.**

I. Introduction

According to People's Republic of China communication industry standard YD/T 1311-2004 technical requirements for the prevention of internet junk e-mail, spam refers to the recipient without prior request or agree to receive advertising email, electronic publications, publicity, and hide the sender's identity, address or contain false information source, sender, routing information, such as e-mail messages. As spammers continue to adopt new methods to make the current mail filtering system becomes inadequate, and therefore the development of safe, reliable spam filter has become one of current problems to be solved. Chinese internet users receive a weekly volume of spam is 13.8, spam accounted for 40.1% [1]. Foreign media reports [2], statistics from security vendor Symantec, as of October 2009, the global volume of spam accounted for 86% of all email volume. Therefore spam filtering research people's wide concern. Mail filter is a very early application, but also a wide range of technologies. Obviously, research more effective spam filtering system, is a significant issue.

Structure of the mail system in accordance with the role of e-mail filtering can be divided into three categories [3], [4]. 1) MTA (mail transfer agent) filter. MTA filtering is the process of the data in the session to check the mail for the line filter to filter processing. E-mail before sending, check for the envelope. E-mail sent later, to inspect the data, including the header check and the letter body check. The most popular message body check is to check the contents using Bayesian algorithm. 2) MDA (mail delivery agent) filter. MDA is a filter to receive mail from the MTA, local or remote to be checked when submitted to deal with line filter mail. Many MDA filters are supported in this process, such as Procmail, Maildrop, and Cyrus-IMAP. 3) MUA (mail user agent) filter. MTA and MDA are server-side email filtering, and MUA filtering is the client-side filtering of mail users. Most popular mail clients such as Outlook Express, Foxmail, have supported the MUA filters. Currently most of the users often enter the mail server to send and receive mail, without the client-side, so the MUA filtering does not work.

There is already hardware spam filters on the market, such as Barracuda's spam filter, its function more perfect. Users can specify their own spam keywords and the extension of the attachment, and can set Bayesian factor to control the degree of intercepting mail. It also has some function, such as anti-virus, automatic-update virus database, but the message content filtering is also not accurate enough.

According to the different stages of email transmission and methods, this paper puts forward intelligent multi-level mail filtering system based on natural language processing. Its main features include black and white list filter, reverse DNS testing, fingerprint recognition, intent detection, content-based spam filtering, and user-personalized filtering technology.

II. Mail-Filtering Related Research

June 1973, J. White submitted a report entitled "A Proposed Mail Protocol" of the RFC, number 524. August 1982, Jonathan B. Postel proposed simple mail transfer protocol (SMTP) based on the famous RFC524. It is defined by the IETF as an international standard. Because at the time not fully taking into account the security of computer networks, there are a lot of security vulnerabilities. Many spammers take advantage of SMTP's security vulnerabilities, such as fake e-mail address of the head or disguise the sender, madly to send spam. It is difficult to identify the real sender.

Manuscript received May 8, 2013; revised October 28, 2013. This work was supported in part by the Ministry of Education Fund (11YJC870011), the Ministry of Science and Technology Fund (2012BAH08B02), Beijing Education Commission Fund (KM201210772014), (61370139), and 2012JGZD07.

Haiyan Kang is with School of Information and Management, Beijing Information Science and Technology University, Beijing 100192, China. (e-mail: kanghaiyan@126.com).

Xiaojiao Yuan is with Computer School, Beijing Information Science and Technology University, Beijing 100192, China.

Although security for SMTP subsequently made a number of enhanced vulnerability agreement, it is difficult to be widely adopted [5]. Existing anti-spam technologies, there are three main directions as following. 1) Modify an existing SMTP protocol to develop a new secure mail protocol, so that spam does not "survival of the environment." The method is through improving communication protocol, to enhance security authentication performance, to eliminate spam flooding the environment, to reduce or eliminate spam. 2) To enable the spammers to bear "a huge cost." Because of the low cost for sending e-mail messages, quickly and easily, spammers send mass advertising information via e-mail. 3) According to the message format, delivery time, file size, content and other characteristics to identify whether the message is spam. This method is the use of message identification technologies to identify and filter spam from messages. Reference [6] proposes a Chinese mail-filtering method with combination of rule-based and statistics-based. Reference [7] proposes a classifying method of Chinese mail based on commixing model of least risking Bayesian.

III. Multi-Level Intelligent Mail-filter Design and the Key Technologies Based on the Natural Language Processing

This research is mainly from a technology perspective, analyze and solve the problem of spam. Principle of content-based mail filtering is to use the keyword technology to block spam. Mail filters are generally connected in front of the server containing the SMTP service, and filtered mail is sent to the specified mail Server. The administrator can select the processing method for not-trusted messages.

The structure of multi-level intelligent mail-filters based on natural language processing, is shown in Fig. 1, which includes black and white list filter, reverse DNS testing, fingerprint recognition technology, intent detection technology, content-based mail filtering, and user-personalized filtering technology. And it can be self learning. content-based mail filtering processing mainly refer to the message subject, body, attachments, including the text content and image content, the study focus on text content filtering.

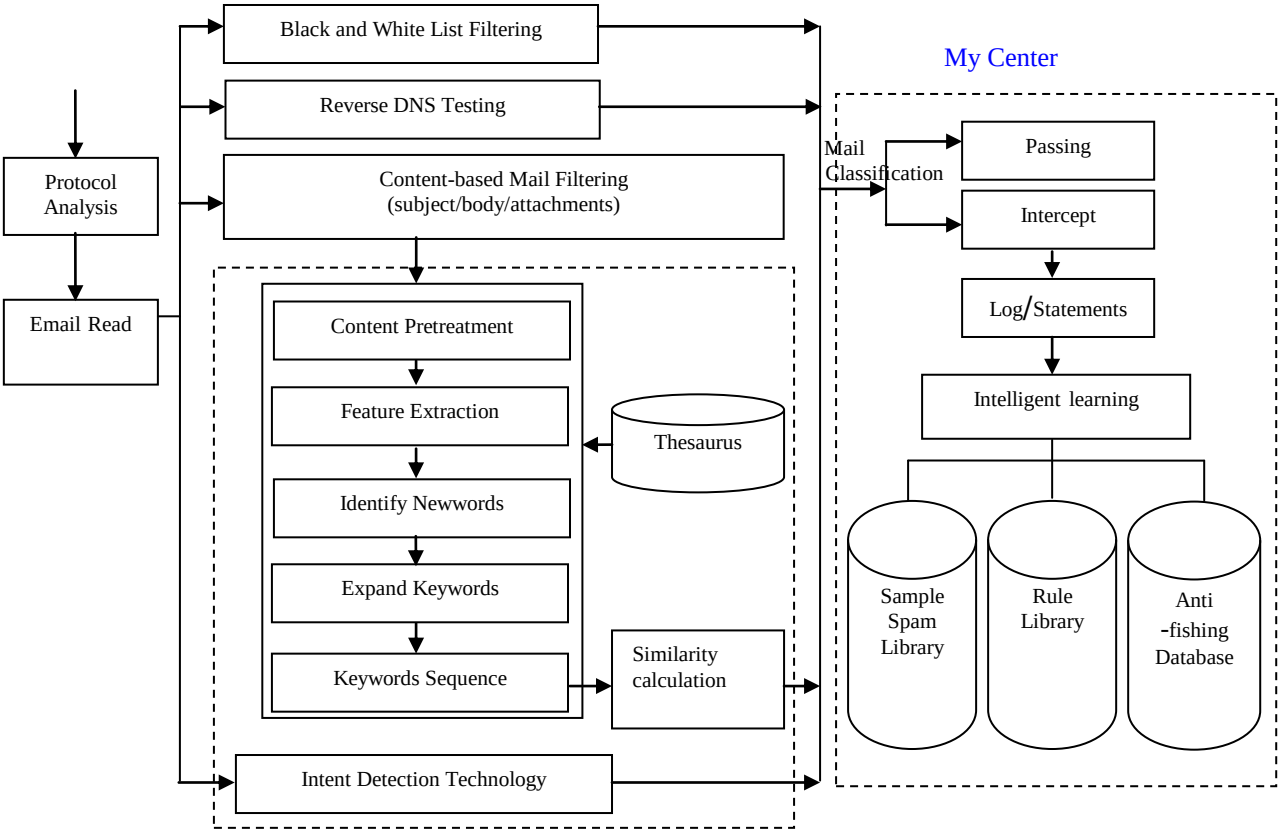

Fig. 1. The structure of multi-level intelligent mail-filters based on natural language processing.

A. Black and White List Filtering

Using an email address of blacklist/white list, the system can block or allow sender addresses directly to enter the enterprise network and arrives in a user's Inbox. Among them, the IP blacklist is a common of spam filtering technology. Manually maintaining a list of spammers IP address, the system filters out all spam messages that are sent by these IP list.

B. Reverse DNS Testing

Reverse DNS is an advanced spam filtering technology. After receiving a message, the system finds the domain of the sender address. If the domain is not a valid DNS MX, the systems determine the address is invalid, and the message will be classified as spam. The return e-mail address can also use the same technology to look accordingly. Meanwhile, the source IP address of received messages can be used to reverse DNS lookup. If the reverse DNS lookup on the message provided the source domain and IP address match, the message is accepted, otherwise rejected.

C. Fingerprint Identification Technology

The so-called mail-fingerprint is some combinations of

string in the message content, also known as a snapshot. It has been recognized as identifying spam messages from similar but not identical information. Spam often contains the following words: proxy services, recruit students, and cash. In fact, this is spam fingerprints, and anti-virus signature recognition technology thoughts are common. The anti-spam filter uses fingerprint confirmation to complete the identification of spam. Of course, it depends on the accuracy of the fingerprint database. Anti-spam products give each character appearing a given value (the value determined in accordance with the specific wording of the law of the characteristics of waste classification), then use statistical methods to calculate this message a comprehensive value. It can also judge on base of many times received messages similarity (received several similar mail is probably spam). The disadvantage of fingerprint recognition technology is to always maintain the fingerprint database.

D. Content-based Mail Filtering

Content-based mail filtering (the system mainly discusses text format), you need to decrypt the message subject, body, and attachments. Then use natural language processing technology to express the message content (including attachments) by the semantic characterization. In this process, first, it need pre-treating the message content, including segmentation, feature extraction etc. Then, it need expand the characteristics items combining with knowledge database, form the vector and match the corresponding categories of the vector database to calculate the similarity of two vectors according to the weight vector, in order to judge the credibility of the message. Furthermore, the system analyses trust-email and non-trust email by self-learning feature, and adjust intelligent the weight of key words related topics, making higher accuracy of spam filters.

1) Intent detection technology

There are a lot of spam, whose title and message body are the same to non-spam, but the message body contains a URL, the URL address of the link is precisely a spam. Intent detection technology is to check the URL and determine whether the message is spam by detecting the link content. The advantage of this technology is to improve spam recognition rate; disadvantage is to regularly maintain illegal URL library.

2) Mail attachments of the recognition and filtering

This project is mainly to read the EML message attachments, and to further confirm its type, judge whether to cheat (if extension name of the attachment is modified). To do the following steps.

"filename="=?gb2312?B?0MK9qM7Esb7OxLW1LnR4dA==?="", this sentence is BASE64 encoded. No matter what type of file attachment "filename =" =? Gb2312? B?" is the same. Which "0MK9qM7Esb7OxLW1LnR4dA ==" represents the attachment name and type. Interpreting the code, you can see "attachment name, attachment type". If the attachment type is "exe", it is very likely that attachment contains a virus. Then confirming with the virus, you need intercept only the attachment to reduce the risk of infectious disease. Then after clipping attachment, it continues to send the message body.

3) Segmentation, keyword understanding and expansion

Chinese is not different with the English which have space between words, so we need to segment question sentence into words for getting the word information of questions. The lexical analysis technology has presently been mature. The accuracy rate of many segmentation programs can be more than 95 percent. We can use the CTCLAS which developed by Chinese Academy of Sciences or the IRLAS which developed by Harbin Industry University. IRLAS adopts the full segmentation method, it matches the sentences in accordance with the word length down, then finds out all possible words and adds every possible word to segmentation word graph. If there is ambiguity segmentation, after the all segmentation, the word graph is a structure graph which has multiple paths. Other methods are also used for segmentation, such as forward maximum matching algorithm, reverse-order maximum matching algorithm, word frequency counting, finite multi-level list, adjacency constraints, association-back, expert system, neural network and so on. This paper segment words with IRLAS.

Keyword understanding: The first step is stop words removing. Stop words are high frequency words in text and their meaning is very poor, which mainly include adverbs, function words, conjunctions, statement label designator and so on. such as "是", "的", "啊", "呀". It will affect the speed if we didn't filter stop words, because of the high frequency of stop words. We are used to filter stop words after lexical analysis in order to enhance the accuracy of question sentences. Stop words database is a collection of meaningless words, we use it to remove stop words. All words will be filtered out as long as they appear in that stop words form. This article uses the Harbin industry university's stop words database.

Keyword expansion: There are fewer words in question sentences, but the words contain a wealth of meaning. If we only rely on the keywords which extracted from the questions to retrieval information, some documents without keyword will be ignored. It will affect the recall rate of candidate web page and the accuracy of the answer extraction. The different question type's answers have different characteristic, so keyword expansion according to the different question type. We can expend keywords according to the question type after determining the type of questions. For example, How much is the volume of regular football? The answers should have "cubic meters" or "cubic centimeters", these words can help to search answers and extract accurate answers. The system collected different question sentence and divided into different categories, then extension keywords depend on the answer of different question type. The system also includes database update, if the searching question is a new question type, the question and the answer will be put into FAQ database.

IV. Experiments and Analysis

Based on the above methods and theories, using B/S triple-tier architecture model in Java EE 6 platform [8] we design and develop a mail filtering system. Using Easy mail

3.3 server, each function has been verified. One of the instances, tests were developed in the database keyword in "法轮功" and "台独", the message with the word will be scanned. Attachment type is exe form. It might contain a virus, so to transfer the attachment. Find the message in the user's inbox and open the message, you'll find that attachment is deleted. System successfully cut out the appropriate code, and to transfer the attachment to the specified folder, as shown in Fig. 2.

In order to test the method proposed by this article, we use Chinese spam corpus collected (CSpam) [9] from CAS institute of Computing for experiments. Test data includes test content, the number of tests, and accuracy, which is shown in Table I.

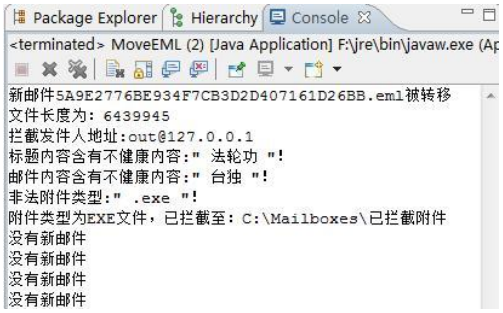

Fig. 2. Exe attachment containing the virus blocked the results.

Large amounts of experiments data show that the combination of multi-level spam filter is feasible and efficient.

TABLE I: EXPERIMENTS DATA OF MAIL FILTER

Test Content	The Number of Tests	Accuracy
User Login	10	100%
Black List Filtering	10	100%
White List Filtering	10	100%
Reverse DNS Testing	50	100%
Fingerprint Identification Technology	50	95%
Intent Detection Technology	200	96.5%
Content-based Mail Filtering	200	90%
Mail Attachments of the Filtering	200	90%
Total	530	96%

V. CONCLUSION

In this paper, a intelligent multi-level mail filtering system was proposed based on natural language processing, and implemented on Java EE 6 platform. Main features of the mail filter include black and white list filter, reverse DNS testing, fingerprint recognition, intent detection technology, content-based spam filtering, and user-personalized filtering technology. Experiments proved that the combination of multi-stage filtration mail filter is feasible and efficient.

In addition, the filter has full message records, intercept mail queries, intelligent interactive features for learning and feedback. This filter is very good possibilities of preventing bypass sending, good scalability to meet enterprise needs. It can also provide anti-virus functionality according to future demand.

ACKNOWLEDGEMENTS

In project implementation and test special thanks to my students Zhang jingwen, Lan ting, Ma aili, Tian liyun, Song shiwei. This project is supported by the Ministry of Education Fund (11YJC870011), the Ministry of Science and Technology Fund (2012BAH08B02), Beijing Education Commission Fund (KM201210772014), (61370139), and 2012JGZD07.

REFERENCES

[1] Anti-spam Center of ISC. Annual Report Anti – spam. (May 24, 2011). [Online]. Available: http://www.anti-spam.cn/

[2] Anti-spam technology and defense. [Online]. Available: http://www.mails7.com/index.php?_m=mod_article&_a=article_content&article_id=160

[3] W. Pan, *A Survey of Content-based Anti-spam Email Filtering*, Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2004.

[4] B. Wang. (July 5, 2007). Chinese spam corpus Cspam. [Online]. Available: http://www.nlp.org.cn/categories/default.php?cat_id=17

[5] W. N. Gansterer, A. G. K. Janecek, and P. Lechner, "A reliable component-based architecture for Email filtering," in *Proc. the 2nd International Conference on Availability, Reliability and Security*, Washington DC: IEEE Computer Society, 2007, pp. 43-50.

[6] J. Wang and J. Zhu, "Design and implementation of chinese-spam filtering system based on Linux," *Journal of Anhui Agricultural University*, vol. 38, no. 2, pp. 309-314, 2011.

[7] N. Wang, J. Zhang, and Y. He, "Algorithm of chinese mail classification based on improved bayesian model," *Computer Engineering and Applications*, vol. 16, no. 3, pp. 57-62, 2006.

[8] X. Yuan and H. Kang, "An algorithm for anonymization of query log," *Journal of Beijing Information Science and Technology University*, vol. 28, no. 5, pp. 24-31, 2013.

[9] H. Kang, Y. Zhang, and W. Liu, "Study on key technologies of generator of Q/A system," in *Proc. IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, December 19-20, 2008, pp. 522-527.

Haiyan Kang was born in China in 1971. He received his Ph.D. in computer application technology from Beijing Institute of Technology, China in 2005. His research interest fields include information system security, natural language processing (NLP). He is currently working as a chief, associate professor at Department of Information Security, School of Information and Management, Beijing Information Science and Technology University, China.

Xiaojiao Yuan was born in Shanxi, China in 1987. She is pursuing master degree from Beijing Infromation Science and Technology University, China. Her major is computer science and technology. Her research interest fields include Information security and natural language processing,and taking part in project and subject of differential privacy.