Web Service for Spam Removal Using Composite Intelligent Algorithm

K. S. Chandrasekaran[1], D. Bhavani[2], S. Divya[3], S. Janani[4], T. Deepa[5]
[1, 2, 3, 4, 5] Saranathan college of engineering, Tiruchirapalli-620002,Tamilnadu,India

Abstract- *An E-mail inbox is filled with both useful and useless mails which are named as "HAM" and "SPAM". It's difficult to maintain and monitor the inbox frequently so the Spam mail may increase day by day. It's difficult to filter those mails manually. The existing project had proposed the logic of spam filtering techniques using composite intelligent algorithm. It is not reliable and is inefficient to use a single algorithm to separate out spam. In order to improve the accuracy and efficiency of spam filtering, composite intelligent algorithm, the proposed algorithm integrates and improves the existing algorithms by utilizing the advantages of previous algorithms and avoiding their shortages. The previous algorithm uses the Rule Configuration, Black listing method and Bayes algorithm. In this paper we convert this logic into "SAAS" (i.e. Software as a Service).Moreover, an intelligent method has the ability of self-learning by using the contents of the e-mails is introduced. Finally, the proposed project shows that the intelligent method with web service achieves a better efficiency, performance and provides more flexibility to allow users.*

Keywords- spam filtering, blacklisting, bayes.

I. INTRODUCTION

With the development of the internet, as a popular way of communication, e-mail is becoming more and more important. Due to the characteristics of lower cost, simple apply, and fast spreading, a lot of unwilling spams appeared on the internet. These spams occupy a mass of bandwidth. It pulls in e-mail's end-user to spend a great deal of time to dealing with them. It is a great challenge for the users who uses e-mail. In order to resolve it, spam intelligent analysis, automatic filtering has already been in development over a few years. Particularly some outstanding technical have appeared in recent years. For example, rule configuration, blacklisting method and statistical theory are often used for spam filtering. But the results of spam filtering are not good when we use a simple filtering technology, because spam has the characteristic of occurrence of variation and analysis difficulty of the e-mail content. A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox. This method is not especially effective, too often omitting perfectly legitimate messages (these are called false positives) and letting actual spam through. More sophisticated programs, such as Bayesian filters or other heuristics filters, attempt to identify spam through suspicious word patterns or word frequency. A spam filter is an email service feature designed to block spam from a user's inbox. Because a large amount of global email messages are spam, effective spam filters are critical to maintaining clean and spam-free inboxes. [1]. Nowadays, combination of various algorithms and techniques occurs in order to improve the efficiency of anti-spam [2]. This paper focus on the research of some kinds of filtering algorithms and revise some disadvantages to achieve a multilayer filtering via an addition method, and by the internal automatic transport of auxiliary to make the algorithm more intelligent and accurate. The paper tries to improve an ability of auto get knowledge by an analysis of e-mail content. And a comprehensive method of spam filtering is proposed. Finally, a comparison between this paper's algorithm and traditional single algorithms will be illustrated through a simulate experiments.

A Web service is a service offered by an electronic device to another electronic device, communicating with each other via the World Wide Web. In a web service, web technology such as the HTTP protocol, originally designed for human-to-machine communication, is utilized for machine-to-machine communication, more specifically for transferring machine readable file formats such as XML and JSON.

II. BLACKLISTING METHOD

It is generally believed that the IP addresses space of spam generally is relatively fixed and regularly [5-7]. The basic idea of the blacklisting method is that some malicious mail senders and suspicious IP addresses will be stored into a database [8]. However, the blacklisting method becomes invalid when mail sender and IP address is changed. As a common algorithm of spam filtering, when a mail arrived in the filtering system, the mail sender in mail header will be collected and compared with black list. If the mail sender is found in black list, a deny action will be triggered. Reversely, if the mail sender is found in white list, a receive action will be triggered. The advantage of this algorithm is occupying less system resources. Nevertheless, there are main two disadvantaged aspects in blacklisting method. Firstly, the contents of black list and white list are pretty accurate. If friendly address listed in the blacklist, the method will cause a false- positive error. It is for this reason that the blacklisting method cannot cover all situations. Secondly, the black list and the white list need to be updated day to day. In a few words, this algorithm is too smart to predict unknown spam mail’s attack.

III. THE DESIGN OF COMPOSITE INTELLIGENT ALGORITHM

The purpose of the composite intelligent algorithm is to address the shortages of the existing algorithms. The basic principle of the algorithm is to identify spam and legitimate email in an algorithm to accurately as possible. The algorithm has the characteristic of trying to learn more and more user’s request and habit, and to reduce the operations of the configuration filter system of e-mail. The composite intelligent algorithm adopts hierarchical filtering architecture. The spam with obvious features is judged by black list method. The legitimate email is judged by white list method. This strategy reduces the intermediate steps of e-mail identifications. The composite intelligent algorithm provides self define rule to filter some e-mails. The core of self define rule is Bayes algorithm and center distance vector algorithm. The identification process of spam is the learning process of e-mail filter algorithms. The learning process includes spam identifications using black list method, legitimate email identifications using write list method, and the learning of self define rule. Black list method is a sort of conservative algorithm. Therefore the database of black list is configured manually. In this way, some legitimate mails will configuration is not necessary for the email filter system. This configuration is only very convenient for users.

Figure 1. Illustrates the structure of the composite intelligent algorithm. The composite intelligent algorithm has 4 main functions: mail filter, mail word classification, mail learning, and mail settings.

3.1 Mail Filter Function

When a mail arrived in the mail filtering system, a mail address or IP address will be collected at the first step. Then the system verifies whether the emails address and IP address is existed in black list. It is generally believed that identifications by addresses are reliable, because vast IP addresses point to the public email system. Therefore, this judgment can lead to misjudgment. Generally, the blacklist method should be carefully used. It is probably to lead a false positive action. In the proposed algorithm, the default blacklist setting is a manually collection. Of course, an automatic black list collection is another option. The extraction of black list depends on the final threshold of algorithm. If the email does not list in the black list, then check the white list if the mail is listed in the white list, mail will be accepted. At the same time, some words are extracted in order to the learning of Bayes algorithm and center distance vector algorithm.

3.2. E-mail Configuration

As can be seen from the chart 1, user can easily check the legitimate mail and spam, and extract the IP address from e-mail header to store into the black list and white list. If user wants to block some legitimate mails due to some personal reasons, the user would add the e-mail address into black list. Equally, if user hopes to receive some spam even it’s really spam, user would add the e-mail address into the white list. Regarding to mail that is sent, the system can automatically extract e-mail address. These email addresses are added to the white list. At the same time, the system can learn the behavior of user through extracting some key words from e-mails. In general, if a user sends out a good mail, the recipient must be a good user, and also the sender could be regarded as a real user. Besides, if the recipient replies a mail to the sender, the IP address of mail reply will be recorded. There are two benefits of this approach.

a) White list can collect the normal IP address without user’s operations.

b) White list has a high reputation; its correctness cannot be impacted by filtering system. Meanwhile, delivered mail is regarded as the main source of Bayes algorithm in the process of the learning of normal mail. The e-mail content that the user receives should closely resemble the email content that the user sends, and that the content of email normally has the similar habit of language using. Furthermore, the return mail that attaches an original message is an important resource on improving Bayes algorithm and center distance vector algorithm.

Intelligent Learning Functions

There are two main algorithms in the proposed system. One is Bayes vector algorithm. Bayes algorithm is used to filter the English characters email, and the center distance vector algorithm is used to filter the Chinese character’s email. The purpose of that is to give full play to the advantages of each algorithm. The Bayes algorithm has been verified to perform well in English environment. The classification performance and precision of the center distance vector algorithm is better than Bayes algorithm. As is shown from the Figure 1, there are two main learning sources in intelligent learning algorithm.

a) The first source comes from the automatically extraction in filter system through some mails such as the spam filtered by the rule algorithm.
b) The second source comes from the manual extraction. Manual extract is not necessary. The purpose of manual extraction only provides an entrance to optimize the system's accuracy and efficiency because a mass of e-mail system asks for a manual input for the materials; the usability of e-mail system is reduced. In this algorithm, with a combination of rule sets, white list, black list, Bayes algorithm and center distance vector algorithm not only improves the accuracy and efficiency of email filtering, but also reduces the burden of user operations, and improves the flexibility of email filtering in email configuration function as well.

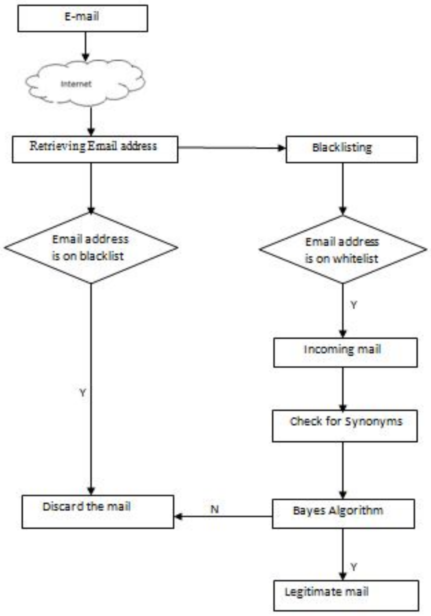

Figure 1

IV. CONCLUSION

In this paper, an algorithm of composite dual engine of spam mail filtering is proposed. From the result of the experiment, the new algorithm is better than any traditional algorithms. It possesses a highest feasibility, greatest automation and intelligence in the whole algorithms. Even so, there still are some rules for improvement in the algorithm such as a collection of signatures in the dictionary. What is more, a threshold of classification value in Bayes needs to be done with deeply research in the future.

REFERENCES

[1] J. Konrad, K. Bartosz and W. Michal, “Application of adaptive splitting and selection classifier to the spam filtering problem", cybernetics and systems, vol. 44, no. 6-7, (2013) October, pp. 569-588.

[2] M. Prilepok, T. Jezowicz, J. Platos and V. Snasel, "Spam Detection Using Data Compression and PSO", Proceedings of 4th International Conference on Computational Aspects of Social Networks, (2012) November 21-23, Sao, Carlos.

[3] N. Pérez-Díaz, D. Ruano-Ordas, F. Fdez-Riverola and J. R.Méndez, “Wirebrush4SPAM: A novel framework for improving efficiency on spam filtering services”, Software - Practice and Experience, vol. 43, no. 11, (2013) November, pp. 1299-1318.

[4] D. Ruano-Ordas, J. Fdez-Glez and F. Fdez-Riverola, “Effective scheduling strategies for boosting performance on rule-based spam filtering frameworks”, Journal of systems and software, vol. 86, no. 12, (2013) December, pp. 3151-3161.

[5] C. M. M. Giovane, S. Anna and S. Ramin, “Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection”, Proceedings of 13th IFIP/IEEE International Symposium on Integrated Network Management, (2013) May 27-31, Ghent, Belgium.