International Journal of Security and Its Applications
Vol.8, No.4 (2014), pp.67-76
http://dx.doi.org/10.14257/ijsia.2014.8.4.07

A Composite Intelligent Method for Spam Filtering

Jun Liu [1*], Shuyu Chen [2], Kai Liu [1] and Yong Zhou [1]

[1]*College of Computer Science, Chongqing University, Chongqing, China*
[2] *College of Software Engineering, Chongqing University, Chongqing, China*
liujuncqcs@163.com, netmobilab@cqu.edu.cn, ielts_kane @126.com, 13545887@qq.com

Abstract

This paper analyses several common algorithms for spam filtering and shows the advantages and disadvantages of these algorithms for spam filtering. Each algorithm is only suitable for filtering specific spam. Some algorithms are suitable for Chinese, and some algorithms perform well in English. In a lot of spam, it is not reliable and inefficiency to using a single algorithm to separate out spam. Thereby, in order to improve the accuracy and efficiency of spam filtering, composite intelligent algorithm, which integrates and improves the existing algorithms by utilizing the advantages of previous algorithms and avoiding their shortages, is proposed. Moreover, an intelligent method that it has the ability of self-learning by using the contents of the e-mails is introduced. Finally, the outcome of experiment shows that the intelligent method achieves a better efficiency and performance.

Keywords: *spam filtering, blacklisting, Rules, Bayes*

1. 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 and stay in a complicated Chinese environment, because spam has the characteristic of occurrence of variation and analysis difficulty of the e-mail content [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. It could fit an intersectional Chinese and English environment. Finally, a comparison between this paper's algorithm and traditional single algorithms will be illustrated through a simulate experiments.

ISSN: 1738-9976 IJSIA

2. Contrast of Common Filtering Algorithm

2.1. Filtering Algorithm via Rule Configuration

The filtering algorithm via configurable rules is known as a heuristic algorithm. It is generally believed that these configurable rules are designed for special spam filters by assembling different anti-spam schemes [3, 4]. It is widely used in the early phase of development. The theory is compared the predefined rules to justify whether a spam mail is. Usually, some keywords, "free, discount, special price" are regarded as the evidence of spam verification. So, the theory depends on the changing demands of customers. The rule of spam mail filtering is customized and maintained continuously.

The algorithm can meet user requirement, such as dealing with the mail header and dealing with mail body. The judgment of two values is the nature of the algorithm. The judgment has many deficiencies, such as dealing with email in two-dimensional space, lack of reliable knowledge, the processing rules of email defined by the user themselves, and very inconvenient use for the user. Moreover, the algorithm requires a lot of time to custom the rules for e-mail. The custom method can not take into account the potential problems.

2.2. 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 can not 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.

2.3. A Intellectual Learning Method Based on the Statistical Theory

There is a classic algorithm called Bayes algorithm, which is used to resolve the classification problems of e-mails. The algorithm exhibits good performance on small data set [9], and has the advantage of running very fast [10]. The core of Bayes algorithm is an application merged with theory of sets and probability statistics.

The idea of the algorithm is as follows:

L is a set of email class. L can be denoted as: $\{l_1, l_2, \ldots, l_n\}$. l_i refers to a kind of mail such as spam, legitimate mail, and privacy mail.

E represents an email, which is a set including some words and phrases. E can be denoted as: $\{w_1, w_2, \ldots, w_n\}$. W_i refers to words and phrases.

$$l^* = \mathrm{MAX}_{lj}\{\Pr(l_j \mid E\} = \mathrm{MAX}_{lj}\{\Pr(l_j \mid w_1, w_2, \ldots, w_n)\} = \mathrm{MAX}_{lj}\{\Pr(l_j)\Pr(w_1, w_2, \ldots, w_n \mid l_j) / \Pr(w_1, w_2, \ldots, w_n)\} \quad (1)$$

l^* represents the probability that mail E belongs to a class email l_j. l^* is the maximum value in $\Pr(l_j \mid E)$. The algorithm is to calculate the maximum value in $\Pr(l_j \mid E)$. Therefore, the value of $\Pr(w_1, w_2, ..., w_n)$ can be ignored. The formula (1) can be simple as:

$$l^* = \text{MAX}_{lj} \{\Pr(l_j)\Pr(w_1, w_2, ..., w_n \mid l_j)\}. \quad (2)$$

With regard to $E = \{w_1, w_2, ..., w_n\}$, there are many possible values in general. If the most contiguous values of Pr (w_1, w_2 ,…w_n | l_j) need to be calculated, a mountain of e-mail experiments must be conducted. To reduce the number of experiments and enhance the reliability of estimated value, Bayes algorithm presents an assumption on the condition independence. For example, assume that every character is conditional independent of other characters in an email class l_i, and considering the location of some characters are very important for the beginning or ending of the text. Obviously, this assumption can not match the real environment. But, Bayes algorithm has applied to various text classification scenarios successfully. According to the assumption of condition independence, formula (2) can be denoted as:

$$l^* = MAX_{l_j}\{\Pr(l_j)\prod_{i=1}^{n}\Pr(w_1 \mid l_j, w_2, ..., w_n)\}$$

$$= MAX_{l_j}\{\Pr(l_j)\prod_{i=1}^{n}\Pr(w_i \mid l_j)\} \quad (3)$$

Pr(lj) is calculated by the ratio of the number of emails belonging to l_j and total number of emails. Conditional probability *Pr(w_1| l_j)* is to calculate the probability of w_1 belonging to some kind of email class in the conditions of that the value of l_j is confirmed. The formula can be denoted as:

$$\Pr(w_1 \mid l_j) = \{1 + N(w_i, l_j)\} / \{T + N(l_j)\} \quad (4)$$

N(w_i,l_j) refers to the number of l_j in W_i. *N (l_j)* is the total number of training set characters coming from the class l_j. *T* is the number of words coming from the training set of the class l_j. In order to avoid the situation with probability 0, the formula (4) uses the Laplace smoothing to add 1 to the number of training set characters.

Start from the original algorithm's thinking, it can be seen that the algorithm is based on context in English (English divides the word with a space). It founded by P. Graham and developed by Arc language. Up to now, the mainstream scenarios of applications are still an English context.

2.4. An Intelligent Algorithm Based on Vector Space

The statistical method does not depend on a particular language context, but only has an ability of identifying spam mail and legitimate mail [11-13]. It is difficult to classify using the statistical method. In contrast, the vector space model, which is one of significant mathematics tools, is better than the statistical method.

The classification algorithm based on center distance is widely used algorithm in vector space model. The algorithm uses a vector space model to represent every data item. Therefore, an email can be regarded as a vector, which can be expressed with a frequency of words appearance in an e-mail. The vector can be denoted as: $\vec{E}_{[tf]} = \{tf_1, tf_2, ... tf_n\}$. tf_i is the frequency that the number of word *i* appears in an e-mail. In this model, an abstract method is used widely. The abstract method checks the frequency of each word in some e-mails because some words appeared frequently in some emails can reduce the accuracy of an email.

Therefore, $tf\text{-}df_{(i)}$ representing some part of an email is introduced, $tf\_df_{(i)} = tf_i * log(N/df_i)$, df_i is the number of the e-mails which include word *i*. *N* refers to the total number of the email. The process is to deal with different lengths of the mail.

In the models, the similarity of two different mails can be judged by the following formula:

$$Cos\ (\vec{e}_i, \vec{e}_j) = (\vec{e}_i * \vec{e}_j) / (\| e_i \|_2 * \| e_j \|_2) \quad (5)$$

Due to the mail is a unit of length, so the formula (5) can be transformed into:

$$Cos\ (\vec{e}_i, \vec{e}_j) = (\vec{e}_i * \vec{e}_j) \quad (6)$$

Assume that the e-mail vector and every vector of an e-mail is confirmed, the center distance vector *C* can be expressed as:

$$\vec{C} = (1 / |E|) \sum_{e \in E} \vec{e} \quad (7)$$

The central vector represents some kinds of mails such as spam or legitimate mail. When a mail come into the filtering system, a comparison based on center distance between spam and legal will be triggered, and will use cosine functions to declare a classification.

The existing problem of this algorithm is that high frequency words can not accurately represent the mail. Limited the range of *dfi* is an effective method for this problem. If the value of *dfi* is greater than specified value, *dfi* is not regarded as decision condition.

2.5. Remains Issue

From the above, it is quite clear that every algorithm has a limitation. In this paper, a composite intelligent algorithm that integrates the advantages of the existing algorithms is proposed. The proposed algorithm tries it best to automate some operation, and provides more flexibility to allow users to operate manually.

3. 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 not be intercepted. Of course, the 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 email address and IP address are existed in black list. It is generally believed that identifications by IP address are not reliable,

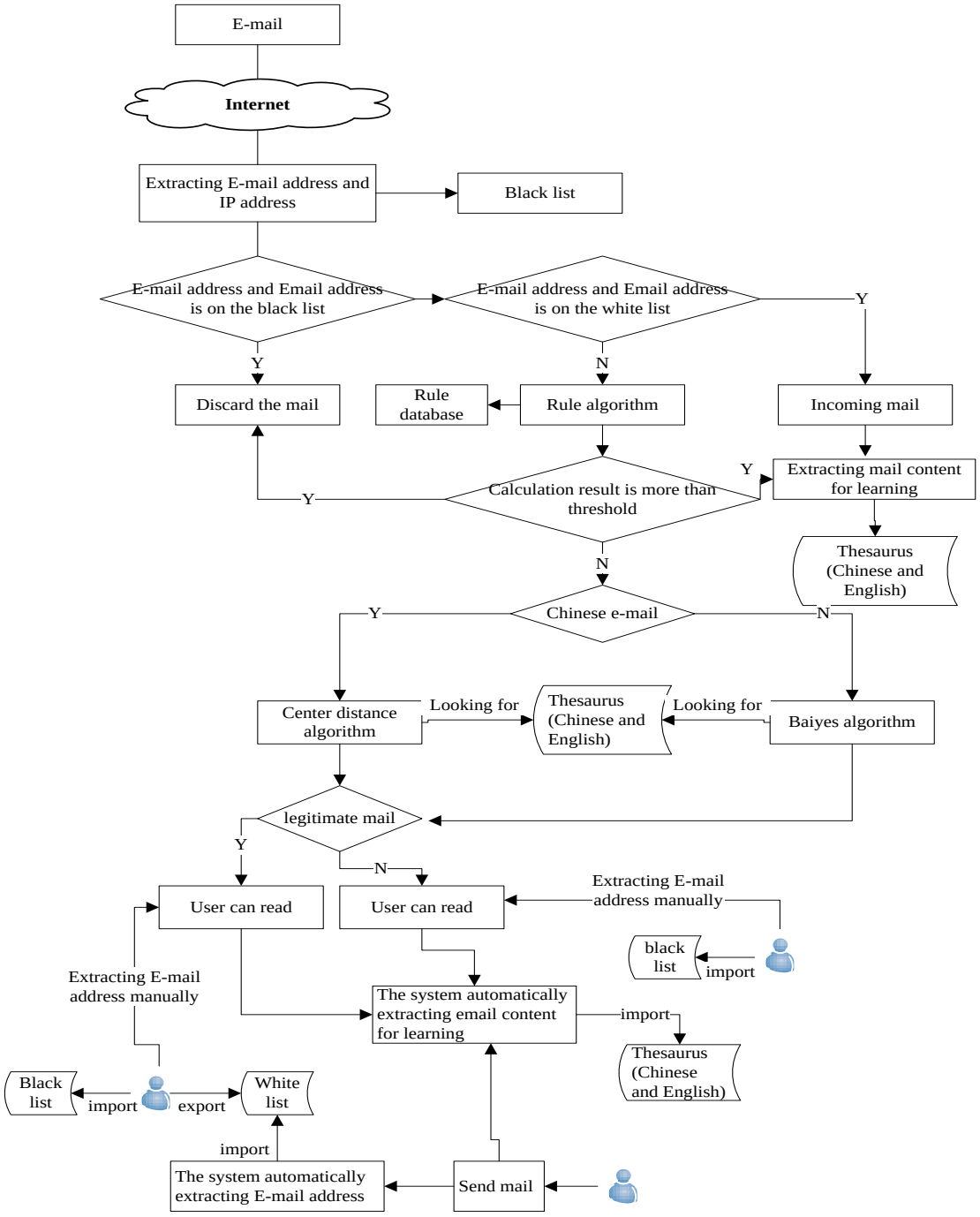


Figure 1. Cache Control Block Maintain Two Caches

because vast IP addresses point to the public e-mail system. Therefore, this judgment can lead to misjudgment. Generally, the black list method should be carefully used. It is probably to lead a false positive action. In the proposed algorithm, the default black list 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. If mail is not listed in black list and white list, it will be judged by regulations algorithm.

The regulation algorithm plays an important role in the mail filter system. In order to achieve a zero false positive goal, the regulation algorithm not only improves the threshold value, but also modifies the value of each regulation that depends on the reality situations. At present, there are some tools to test the effectiveness of these regulations. In this algorithm, the system can automatically test these regulations. The default setting of the proposed system is to check the effectiveness of these regulations periodically. For example, the regulations are used in the identifications of spam mail and legitimate mail. If a regulation matches a number of legitimate mails, the regulation will be deleted and the value of the regulation will be decreased. If some mail regulations are matched by massive spam, the value of these regulations will be increased, because the value of these regulations can not reach the specified threshold, and the degree of match is not high enough in legitimate mail.

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 can not 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.

3.3. Intelligent Learning Functions

There are two main algorithms in the proposed system. One is Bayes algorithm; another is center distance 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, the e-mails which are sent from local, and the e-mails which are accepted using white list. The filter system can be collected valuable information from these e-mails. The important point of the intelligent learning is automation, which means that the user doesn't have to take part in the operation of email filtering.

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.

3.4. Feature Items Selection in Vocabulary

Eigenvector of the file can not include all the words. Therefore, it is necessary to select an efficient algorithm. In this paper, the feature selection algorithm is as follows: firstly, some frequently undistinguished pronouns and adverbs would be diminished. Secondly, the words whose frequencies of appearance less than 3 times by using ZipF rule for both junk and regular emails are removed. ZipF rule means that the probability of the second frequent appearance word is 1/2 of that of the most frequent appearance one. And the probability of the third one is 1/3 of that of the most frequent appearance one. If the occurrence number of the most frequent appearance word is N, the number of second one is $(i/i \times N)$.

4. Experimental Evaluations

In order to investigate the effectiveness of the proposed composite intelligent algorithm, the proposed composite intelligent algorithm is compared with existing algorithms, they are the black list algorithm, the white list algorithm, the rule set algorithm, the Bayes algorithm, and the center distance vector algorithm. A conclusive outcome is drawn. The paper makes samples of 200 regular emails (100 Chinese and 100 English involved) and 200 junk emails (100 Chinese and 100 English included). And some non-functional pronouns and adverbs in each of them are deleted. Assessment indexes of the arithmetic are recall ratio and false positive rate, respectively.

The recall ratio can be denoted as:

Recall ratio = the number of correctly identify mail / the total number of spam.

The false positive rate can be denoted as:

False positive rate = the error number of spam recognition as normal mail / the total number of normal mail.

Table 1. The Comparison of these Algorithms

Performance Comparison	Composite intelligent algorithm	Black list and white list algorithm	Rule algorithm	Baiyes algorithm	Center distance victor algorithm
Recall ratio	95.2%	6.1%	70.5%	80.1%	83.8%
False positive rate	<1.5%	0	0	6.3%	5.6%

Table I shows the experimental results of all kinds of algorithms. Bayes algorithm and center distance vector algorithm both filter 200 emails after learning 200 emails. Composite intelligent algorithm generates outcome automatically. According to the experiment, recall ratio of composite intelligent algorithm is far more than that of others, as 95.2% high. If users take advance of black list to adjust rule algorithm and learning materials manually, performance of recall ratio and false alarm rate would be better.

It is sure that performance of filtering will be improved with development of manually sending email and learning materials.

5. 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 rooms for improvement in the algorithm such as a collection of signatures in the dictionary. What is more, a threshold of classification value in center distance vector needs to be done with deeply research in the future.

Acknowledgements

We are grateful to the editors and anonymous reviewers for their valuable comments on this paper.

The work of this paper is supported by National Natural Science Foundation of China (Grant No. 61272399) and Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20110191110038).

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.

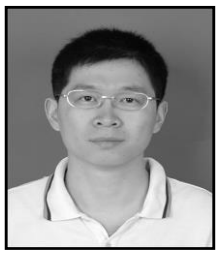
[5] C. M. M. Giovane, S. Anna and S. Ramin, “Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection”, Proceedings of 13th IFIP/IEEE International Symposium on Integrated Network Management, **(2013)** May 27-31, Ghent, Belgium.

[6] M. A. Rajab, F. Monrose and A. Terzis, “On the effectiveness of distributed worm monitoring”, Proceedings of the 14th conference on USENIX Security Symposium, vol. 14, **(2005)** July 31–August 5, pp. 15-15, Berkeley, CA, USA.

[7] M. P. Collins, T. J. Shimeall, S. Faber, J. Janies, R. Weaver, M. De Shon and J. Kadane, “Using Uncleanliness to Predict Future Botnet Addresses”, Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ser. IMC ’07, **(2007)** October 24-26, pp. 93-104, San Diego, USA.

[8] R. Anirudh, F. Nick and V. Santosh, “Filtering spam with behavioral blacklisting”, Proceedings of the 7th ACM Conference on Computer and Communications Security, **(2007)** October 2 – November 2, pp. 342-351, Alexandria, USA.
[9] L. E Zhang, J. Zhu and T. Yao, “An evaluation of statistical spam filtering techniques”, ACM Transactions on Asian Language Information Processing, vol. 3, no. 4, **(2004)** December. pp. 243-269.
[10] X. Ma and Y. Shen, "Combining Naive Bayes and tri-gram language model for spam filtering", Advances in Intelligent and Soft Computing, vol. 123, **(2011)** December, pp. 509-520.
[11] N soonthornphisaj, "Anti-spam filtering: A centroid-based classification approach", Proceedings of the 6th International Conference on Signal Processing and communication systems, **(2012)** December 12-14, Radisson , Australia.
[12] A. Wang, W. Fan, J. Wu and Y. Shi, "Spam filtering system study based on 2v-SVM", Proceedings of the 7th World Congress on Intelligent Control and Automation, **(2008)** June 25-27, pp. 4213-4216, Chongqing, China.
[13] X. Sun, Q. Zhang and Z. Wang, "Using LPP and LS-SVM for spam filtering", International Colloquium on Computing, Communication, Control, and Management, CCCM, vol. 2, **(2009)** August 8-9, pp. 451-454, SanYa, China.

Authors



Jun Liu, he received his B.S. degree in Southwest University, P. R. China, at 2001, and M.S. degree in Chongqing University, P. R. China, at 2009. Currently he is a Ph.D. candidate in College of Computer Science, at Chongqing University. His current interests include flash memory, information security, Linux Kernel and big data analytics.

ShuYu Chen, he received his Ph.D. degree in Chongqing University, P. R. China, at 2001. Currently, he is a professor of College of Software Engineering at Chongqing University. His research interests include embedded Linux system, distributed systems, cloud computing, etc. He has published over 120 journal and conference papers in related research areas during recent years.

Kai Liu, he who got a Bachelor of Science in Computer Science at Southwest University, P. R. China and Master of Science in Informatics at University of Reading, UK, is much familiar with techniques of Information and email Security

Yong Zhou, received his B.S. degree in Chongqing University, P. R. China, at 2002, and M.S. degree in Chongqing University, P. R. China, at 2007. His current interests include date backup and recovery, information security, Linux Kernel.