A Keyword Filters Method for Spam via Maximum Independent Sets

HaiLong Wang[1], FanJun Meng[1], HaiPeng Jia[2], JinHong Cheng[3] and Jiong Xie[3]

[1]*Inner Mongolia Normal University*
[2]*Air Defence Forces Academy*
[3]*Inner Mongolia electric power information and Communication Center*
lzjtuwhl@163.com

Abstract

In order to evade the keyword filtering, the spammers insert comments into e-mails, such as unusual symbols # or ※ to divide some keywords. In the paper, one keyword filters method for spam via maximum independent sets is presented, and the main contents include: (1) build a matching relation matrix algorithm to help us to improve the performance of maximal independent sets; (2) develop a judgmental criterion according to the matching relation matrix algorithm. (3) design a behavior recognition technology, which can detect and reject the email which receiving. Proved by the experiments and analyses of examples, the space and time complexity of this algorithm is much smaller than 0 (mn). The operating efficiency is also satisfactory, and is able to achieve the complete filtering of targeted unusual symbols during the e-mail keyword filtering.

Keywords: *Maximum independent sets; semi-diagonal line; keyword filters; matching relation matrix*

1. Introduction

E-mail has become the important method for network communication as it is widely used among the Internet users and is regarded as one of the most commonly used network applications. However, with the development of Internet, junk e-mails (spam) bothering most people do not only bring discontent to users, but also cause some web security issues and economic losses. In recent years, a large number of dedicated servers generate and send out spam though email. According to the statistics from the anti-spam center of the Internet Society of China, There are more than 100 servers per month which were sent into the blacklisted by authoritative foreign anti-spam organization since 2005. With the development of the network, spam becomes an increasingly serious global security problem. More and more researchers pay attention and concern to this field.

Negative effects derived from spam, which bring great economic losses and result in large amounts of data and information blockage, have become a worldwide problem. Numerous experts and scholars put forward a lot of targeted prevention methods that ease the problem to some extent. Data Mining is a relatively popular technique for the filtering of e-mail contents and theme keywords, which detect spam keywords through keyword classification and statistical algorithm. Bayes filter is an effective method. The characteristics of Bayes filter are adaptation and self-learning. Bayes filter has the advantages of high detection accuracy [1]. Other widely used detecting approaches include detection based on memorial information, detection based on description of event features, and filtering based on spam feature analysis and regular expression matches.

2. Related Filtering Technology and Analysis

In Anti-Spam Solutions and Security, Dr. Neal Krawetz sorts the anti-spam techniques into four main categories: Filter, Reverse lookup, Challenges, and Cryptography. All of these solutions can ease the spam problems, but have respective limitations. Filter is a relatively simple and widely used technique in spam detection. It is commonly used in mail receiving management system for judging and eliminating spam. Currently, most of the mail servers apply anti-spam plug-ins, detection gateways and client-side spam filtering, based on varieties of filtering techniques such as keyword filtering algorithm, black and white list, rule-based filtering and Bayesian filtering algorithm [2].

2.1 Keyword filtering algorithm

Mail content filtering uses a key word or multiple keywords as a general basis to judge. Using keyword hit rate to confirm whether this message is spam or not, if the hit rate is large than a set threshold, it is considered as a spam. In addition, the key words can also be phrases and short sentences. Mail header information is the original record of the mail delivery process which is also a very important sense data, this is spam. Spammers use various tools and randomly send the span to forged closing sender, subject and content, but some common information stored in the message header information, which consist IP address, host name, X-identification [3]. Through the filter of this information, you can find out the spam from several mail sent from the same address which includes different transceiver address and subject.

Keyword filtering algorithm generally created some spam associated keyword table to judge and process spam. If certain keyword appears in large numbers of spam email, then we can put them on the filtered list. The defect of this algorithm are there is a great impact on filtering capacity, the selection procedure costs a lot resources, and relatively low efficiency in the stage of selection of keywords. In addition, the word-split function and word-combination function can easily avoid the filtering.

2.2 White and Black List

The blacklist contains the basic information of confirmed spammers: IP address or IP address of the mail sender. If the sender of the mail is the same as the address of the known spam, we can judge this message as a spam, and reject the mail. The disadvantage of this method is that the can bypass the blacklist detected by using a different IP address. The spammer can use the forgery and changed sender's address [4]. In addition, rely on manual processing administrator is unable to update the blacklist in time effectively.

The whitelist contains the trusted e-mail address, or IP address. If the information of send’s mail matches the data in whitelist. This mail can be considered as a normal mail and was released. The disadvantage of this approach is that if the user wishes to receive e-mail from a certain address, the user receives this address mail rules must be set in advance to allow. If a user want to change his/her mail address, the whitelist should be updated in time. Otherwise, the mail server will reject this mail.

2.3 Rule-based Filtering

The rule-based filtering technology defines the filtering expressions or rules mainly by selecting certion characteristics of keywords to describe spam’s feature value. The fatal defect of this filtering method is that it requires managers to maintain a relatively large rules library,

and in order to keep the effective and real-time of the system, managers need to organize new rules regularly [5].

2.4 The static content filtering technology

Static content filtering actually is only useful for the ruling spam. For example, marketing advertising is a spam which always contains these rules, the advertising mail includes a subject if "ADV". If the user does not want to receive the spam like this type, he can simply set a filter to reject the mail which subject with the "ADV" tag. However, the further mail spam appears, for example, some of the words "free" was transformed into "Free ... fee" or "Free - fee" spam. It becomes another hot problem in filtering fields. If we scan the mail and reject the mail which includes these types' data, some normal mail will be deleted. Therefore, the keyword filtering technology based on the message content will lead to a high false rate in practical. This technology can be used in the environment under highly controlled [6].

2.5 Bayesian Algorithm

Bayesian filter mainly calculates the probability that whether an e-mail includes a spam message content (TOKEN string), trains from the manually identified spam and legitimate mail. Thus the result is more effective than other average content filters. Bayesian filter is a score determined filter which utilize the automatic creation of spam feature table. The algorithm first analysis respective feature value in massive spam and legitimate mails, and then calculate the probability that multiple features contains in the mail. The principle of this algorithm is to check the keywords in spam set and legitimate mail set, statistics each feature value as a TOKEN string, then builds hash tables respectively for TOKEN string in the spam set and legitimate mail set according to the occurrences of the extracted TOKEN string, named as word frequency. And these tables store the mapping relationship from TOKEN string to the word frequency. We compute the probability that TOKEN string exists in the hash tables by $P = \frac{WF}{L}$, WF is the word frequency of a certain TOKEN string, and L is the length of corresponding hash table. P indicates the probability that new received email is spam when the mail content contains certain TOKEN string in the hash tables established by system [7]. Finally, we get the spam decision score from the overall mail, calculating by the composite probability formula we obtain:

$$P(A \mid t_1, t_2, \cdots\cdots t_n) = \frac{P_1 * P_2 * \cdots\cdots P_n}{P_1 * P_2 * \cdots\cdots P_n + (1-P_1) * (1-P_2) \cdots\cdots (1-P_n)} \quad \textbf{(1)}$$

P（A| t_i）, where P_i ($1 \le i \le n$) denotes the probability that a mail is spam when it contains the TOKEN string. If the result is greater than a specified threshold, we set the mail as the spam; otherwise, the mail is legitimate. What we can observe from the Bayesian algorithm that spammers may escape the Bayesian algorithm filtering by random inserting a word or sentence. Because of these filter using the static passive detection technology, many of them can be most effective only within a very short period. In order to maintain the effectiveness and real-time of spam detection, the managers should update the rules of filter constantly [8].

Currently, the main anti-spam systems commonly used keywords filtering technology based on complete matching, intercepting samples, analyzing characteristics, generating rules,

rules issued and content filtering technology. In order to avoid special keywords filtering, spammers often insert a large number of comments into the e-mail in order to split some keywords (such as 法轮功) and mix the mail content by some special methods. for example: inserting symbols into Keywords "法＃轮※功" *etc*. Sometimes they convert promotional content into a Zip package as the attachment to evade the filter.

In this paper, we design and implement an adopted algorithm which can effectively solve the problems including keywords split and combine, as well as inserting special symbols. Considering the traditional filter technology always check the email after all the content download into the local disk, which downgrade the performance. We also design a behavior recognition technology, which can detect and reject the email which receiving. In this way, we do not need to wait all the email content fully received from the remote nodes, and directly block the email at the very beginning of the email transfer. The entire filter rule will be build at the initial period of the establishing SMTP connection.

3. The Keywords Filtering Algorithm Based on the Maximal Independent Set

3.1 Matching relation matrix of string

Given any two strings S and T, the maximum matching problem of them is equivalent to the maximal independent set of matching relation matrix.

Define $S = a_1 a_2 \cdots a_m, T = b_1 b_2 \cdots b_n$.

Note that $< n >= \{1, 2, \cdots, n\}, < m >= \{1, 2, \cdots, m\}$

Thus $\{(i, j) \mid i \in < m >, j \in < n >, a_i = b_j\}$ is called matching relation set of S and T, written $M(S,T)$ [5], here we assume that $n \geq m$ generally [9]. C matching relation matrix, can be defined as follow:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m1} & \cdots & c_{mn} \end{bmatrix} \quad (2)$$

Where $c_{ij} = \begin{cases} 1 & ,if \quad a_i = b_j \\ 0 & ,if \quad a_i \neq b_j \end{cases}$

We only discuss the situation that weight $C_{ij} = 1$ in this paper.

Definition This thesis defines that each node in the independent set only exist a corresponding node at the bottom right located in the different row or different column, called quasi diagonal.

The set of nodes which value are 1 ($C_{ij} = 1$) over a quasi diagonal of matching relation matrix is called an independent set. The keyword matching problem can be transformed into solving maximal independent set of matching relation matrix, and searching all independent sets in a given matching relation matrix. So the longest set is the answer. In particular, we can search points in matching relation matrix that of which value is 1 to determine whether they are completely match in the process. However, the idea discuss above will make the problem more complicated, we find that finding the maximal independent sets can be regarded as searching for a road in the matching relation matrix of which value is 1. Each node in the independent set only exist a corresponding node at the bottom right located in the different row or different column, called quasi diagonal, search the bottom right according to the quasi diagonal [10].

We introduce a new problem solving algorithm to find the independent sets in this paper. We will describe the detail in Section 3.2. The relationship between output result with original input string will satisfy the following relationship:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \quad \textbf{(3)}$$

3.2 The algorithm of maximal independent set of matching relation matrix

In order to search the bottom right corresponding with the quasi diagonal, we propose an improved algorithm of maximal independent set of matching relation matrix as follows:

Assume string α and β are independent sets obtained in searching, we set the length of them as |α| and |β|.

The procedure of algorithm describes as follows:

Step 1. When search the column j, if |α| ≥ |β|, and the abscissa of the last character of α i1 < i2(the abscissa of the last character of β), then stop the operation of β.

Step 2. When search the row i, if |α| ≥ |β|, and the ordinate of the last character of α j1 < j2 (the ordinate of the last character of β), then stop the operation of β.

Step 3. When searching matching relation matrix is completed, the length of α is equal to the original string, then keywords are found [11].

In this algorithm, it may generate multiple result of β, because we just find the length of α with the length of original string, so we can get multiple independent sets in the end. The finding procedure will be present as a pseudo code in following [12]:

```
search
  d[]=0；col=0；η[]=0；k=0；
  for j∈col,…,n do
    for i∈k,…, m do
    if equal(S[i],T[j])
      { d[]++；col++；k++； continue；}
    if(j<n&&k>=m)
      k=0;
  End of For
    η[] = d[]/m；
  End
```

(4)

Where d[] indicates the array of length of each independent set during the storage and calculation, η[] presents the matching accuracy of each independent set, col is the next matching start position of target string T[j], This setting design of col can largely reduce the matching time complexity. K denotes the identity of search the original string. When the first searching trip of the original string is finished and the target string is not completed, and then search from the first character of the original string again, the search process will stop until it finds all the matching string [13].

3.3 Judgement Criterion

The Matching accuracy $\eta = \frac{N\alpha}{N_c}$, where N_α is the length of original keywords string, $N_c = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}$ is the length of all quasi diagonal (Cij only equal 0 or 1). If η<1, indicates the string α and the detecting target string does not match exactly, then the system outputs the results: the mail is secure; if η>1, meaningless; if η = 1, then the system shows the string α and the detecting target string match exactly, keywords hidden in the string β is found, then the system give a warning, and continue the next steps [14].

3.4 Complexity Analysis

In this paper, we also import two techniques in our algorithm. Firstly, the matching relation matrix is created dynamically; we do not need any other saving space to store the matrix table. Secondly, In the procedure of string matching, we does not require all the elements of the original string and the target string equal, we just search along the bottom right of the quasi-diagonal [15], the search detail can be find on the algorithm pseudocode description of definition of col. Both the space and time complexity of this algorithm are far less than O (mn).

4. Example Analysis

Assuming that the word "法轮功" as a keyword in the list of filtering keywords, we detect the received mail subject “法＃轮％功※法*律”, with our algorithm designed in this paper. All the process steps can be describe as follows: firstly, the Chinese matching relation matrix as show below:

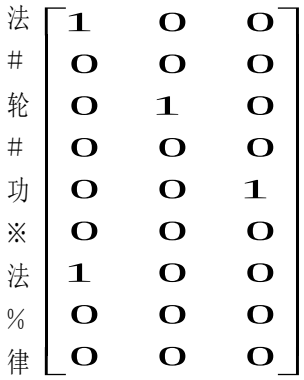

Figure 1. Matching relation matrix (1)

When search along the bottom right of the quasi-diagonal according to matching relation matrix 1, the entry at the row labeled 1 and the column labeled 1 is 1,we firstly get the string α = "法", there is no relevant matching in the second line, then go to the third line, obtain α = "法轮"; similarly and so on, we can get α = "法轮功"; When the program running into the seventh line, the first column is 1, but it is not the last node of string α, so it create a new string β. Here we cannot determine whether the length of β is the same as the keywords, thus continue search β="法". When all the searching procedure finished, we get final string α [8]. We can use the matching accuracy algorithm to determine whether the keyword is a spam, and then carry out the corresponding treatment [16].

To illustrate the effectiveness of the proposed algorithm, the characters matching process via the relation matrix shows in below [17]:

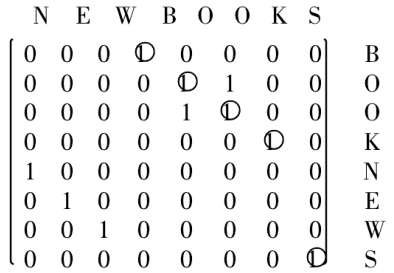

Figure 2. Matching relation matrix (2)

First of all, we get a temporary string ϕ = B (node on ϕ ①). And in the right, we find a way, row 4 contains two ①, we can select the above one accordance to the rules of the algorithm for solving a 1, that is to select the first row 1, instead of 1 in the second row. Meanwhile, in the first row there are two 1, by Theorem 3, we select the left of that one, i.e. the fourth column is a. In this case, φ = BO. However, when the algorithm is running into the fourth row, φ = BOOK, K in the sixth column of the third line of the Bank in the first column, the left side of the road ϕ is the last node of K, then you must create a new one road Ψ, because we cannot determine whether future there | ϕ | ≥ | Ψ |. When the algorithm is run to the sixth row, ϕ = BOOK Ψ = NEW | ϕ | = | Ψ | = 3, we will the S chain on to the road φ, obtained the longest lower right road φ = BOOKS | φ | = 5. Thus, it is possible to calculate the degree of matching of the two strings [18].

5. Conclusions

In this paper, we proposed a keyword filtering methods to filter the spam which is related as the approximate matching of Chinese characters. The basic idea of our design is based on the concept of the edit distance approximate string matching. Edit distance is defined as number how much times a string transform into a minimum number of edits needed another string. By calculating the edit distance matrix, we can draw achieve the best match. Search through the parallel simulation, you can speed up the running process of the classical algorithm; this method is especially good for the short string [19]. If the bit comparing process, we import the parallel principle idea to the matching function, a number of different values packed into a computer word length "w", these words can be processed in a single operation or operator which need several operations or operator to complete the function in traditional methods. To judge a text string in a location or the pattern strings is matched or not, which may match more easily than judgment. If the filtering algorithm cannot successfully match the region, and then combined with the non-the filtering text search algorithm, we can also ultimately achieve fast string matching [20].

This paper presents a set of mail keywords filtering methods to find the maximal independent set. We design and implement an adopted algorithm which can effectively solve the problems including keywords split and combine, as well as inserting special symbols. We also design a behavior recognition technology, which can detect and reject the email which receiving. The experimental result shows that both space and time complexity are far less than O (mn), the efficiency is also satisfactory. However, if the spammers change the produce way and constantly sending the spam, we will always in a passive position. In future, we will further study the anti-spam technology to change our passive reaction position to impassive detecting the spam.

Acknowledgments

This research was supported by Scientific Research Project of Higher Education of Inner Mongolia Autonomous Region, China (NJZY13052).

References

[1] Z. Li and H. Shen, "SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box", INFOCOM, 2011 Proceedings IEEE, **(2011)**, pp. 1835-1843.
[2] M. T. B. Aun, B. -M. Goi and V. T. H. Kim, "Cloud enabled spam filtering services: Challenges and opportunities", Sustainable Utilization and Development in Engineering and Technology (STUDENT), 2011 IEEE Conference on, **(2011), pp. 63-68**.
[3] Q. Luo, B. Liu, J. Yan and Z. He, "Design and Implement a Rule-Based Spam Filtering System Using Neural Network", Computational and Information Sciences (ICCIS), 2011 Int'l Conference on, **(2011)**, pp. 398-401.
[4] P. Graham, "Better Bayesian Filtering", http://paulgraham.com/better.html, **(2003)** January.
[5] C. Dwork and M. Naor, "Pricing via processing or combatting junk mail", in Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology, Springer-Verlag, **(1993)**, pp. 139-147.
[6] A. Li and H. Liu, "Utilizing improved Bayesian algorithm to identify blog comment spam", Robotics and Applications (ISRA), 2012 IEEE Symposium, **(2012)**, pp. 423-426.
[7] A. C. Yao, "The Complexity of Pattern Matching for a Random String", SIAM Journal on Computing, vol. 8, no. 3, **(1979)**, pp. 368-387.
[8] J. A. Bondy and U. S. R. Murty, "Graph Theory with Applications", The MaCmillan Press ltd, London and Basingstoke, **(1976)**.

[9] S. Ghemawat, H. Gobioff and S. Leung, "The Google File System", SIGOPS Oper. Syst. Rev., vol. 37, no. 5, **(2003)**, pp. 29-43.
[10] R. Drewes, "An artificial neural network spam classifier", **(2002)** August, http://www.interstice.com/drewes/cs676/spam-nn/spam-nn.html.
[11] H. Yuan and D. Wang, "The New Approach of Marking Activity-Loops Based on the String Reachable Matrix", Communications and Mobile Computing, 2009, CMC '09, WRI International Conference, **(2009)**, pp. 569-572.
[12] E. Irshad, W. Noshairwan, M. Shafiq, S. Khurram, A. Irshad and M. Usman, "Performance Evaluation Analysis of Group Mobility in Mobile Ad hoc Networks", International Journal of Future Generation Communication and Networking (IJFGCN), Syst. Rev., vol. 3, no. 3, **(2010)**, pp. 33-40.
[13] J. -C. Lin and T. C. Huang, "An efficient fault-containing self-stabilizing algorithm for finding a maximal independent set", Parallel and Distributed Systems, IEEE Transactions, **(2003)**, pp. 742-754.
[14] G. Vesztergombi, G. Odor, F. Rohrbach and G. Varga, "Scalable matrix multiplication algorithm for IRAM architecture machine", Parallel and Distributed Processing, 1998, PDP '98, Proceedings of the Sixth Euromicro Workshop, **(1998)**, pp. 367-372.
[15] C. -H. Lin, J. -C. Liu, C. -T. Kuo, M. -C. Chou and T. -C. Yang, "Safeguard Intranet Using Embedded and Distributed Firewall System", International Journal of Future Generation Communication and Networking (IJFGCN). Syst. Rev., vol. 2, no. 1, **(2009)**, pp. 9-16.
[16] J. Xie, S. Yin, X. Ruan, Z. Ding, Y. Tian, J. Majors, A. Manzanares and X. Qin, "Improving MapReduce performance through data placement in heterogeneous Hadoop clusters", in: Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, doi:10.1109/IPDPSW, **(2010)** April, 5470880, pp. 1-9.
[17] S. -C. Kim and J. -M. Chung, "Message Complexity Analysis of Mobile Ad Hoc Network Address Autoconfiguration Protocols" Mobile Computing, IEEE Transactions, **(2008)**, pp. 358-371.
[18] S. V. Viraktamath and G. V. Attimarad, "Impact of Quantization Matrix on the Performance of JPEG", International Journal of Future Generation Communication and Networking (IJFGCN), Syst. Rev., vol. 4, no. 3, **(2011)**, pp. 107-118.
[19] M. Lacoste, "Architecting Adaptable Security Infrastructures for Pervasive Networks through Components", International Journal of Future Generation Communication and Networking (IJFGCN), Syst. Rev., vol. 3, no. 1, **(2010)**, pp. 1-16.
[20] F. Marozzo, D. Talia and P. Trun-fio, "A peer-to-peer framework for supporting mapreduce applications in dynamic cloud environments", in Nick Antonopoulos and Lee Gillam, editors, Cloud Computing, vol. 0 of Computer Communications and Networks, Springer London, **(2010)**, pp. 113–125.

Authors

Hailong Wang received the BS in computer science from North Jiaotong University, China, in 1998,and received the MS in computer science from Lanzhou Jiaotong University, China,in 2007.Currently,he is an assistant professor in Computer & Information Engineering College at Inner Mongolia Normal University, China. His research interests include embedded system and muti-core processors, and also fault tolerance and real-time database.

Jiong Xie received the BS and MS degrees in computer science from BUAA (Beijing University of Aeronautics and Astronautics), China, in 2004 and 2008. He is currently working toward the PhD degree at the Department of Computer Science and Software Engineering, Auburn University. His research interests include scheduling techniques and parallel algorithms for clusters, and also multi-core processors and software techniques for I/O-intensive applications.