INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2022.16.62 Volume 16, 2022

# The Optimization Analysis of Phishing Email Filtering in Network Fraud based on Improved Bayesian Algorithm

Yahao Zhang, Jin Pang, Hongshan Yin
State Grid Information & Telecommunication Branch, Beijing100761, China

Received: July 27, 2021. Revised: December 28, 2021. Accepted: January 12, 2022. Published: January 13, 2022.

**Abstract— Mail transmission was not only the main function of information system, but also the main way of network virus and Trojan horse transmission, which has a key impact on the running state of information. In order to deal with the threats of network viruses and Trojans and improve the level of e-mail management, this paper studies the filtering of information system, and proposes a phishing e-mail filtering method based on Improved Bayesian model. MATLAB simulation results show that the consistency p between the amount of data sent by e-mail and the amount received is good, the consistency rate reached 92.3%. the data security level** is 95%**, encryption proportion / data proportion ratio under Bayesian optimization are higher than those of unfiltered method，which up to 97.2%. Therefore, the Bayesian optimization model constructed in this paper can meet the needs of phishing email filtering in information communication at this stage.**

**Keywords—Bayesian algorithm, Information system, Phishing email filtering, Fourier series, Hashtable**

## I.INTRODUCTION

With the rapid development of information technology, the implementation of the "information speed" strategy and the proposal of the concept of "global village", the construction of information network has almost doubled from 48000 servers in 2012 to 74300 servers in 2020 [1]. At the same time, the extensive use of WiFi, Internet of things, LAN and mobile technologies has also increased the number of spam, Trojan mail and virus mail in the information system from 371.2 billion in 2012 to 9723 trillion in 2020 [2]. By the end of 2020, the number of users in China has reached 1.172 billion, accounting for 92.3% of China's total population. Under such a huge demand for data processing, the virus prevention bearer of the information system is increasingly overburdened, and even threatens the stability of the operation of the information system, which also puts the "optimization problem of phishing mail filtering" on the agenda [3]. At present, the mail receiving of the information system mainly adopts the hoop topology, which has the advantages of stable, efficient and strong compatibility in data processing, and can also meet the needs of mail complex information processing [4]. However, in the context of a large number of emails, the hoop topology will reduce the filtering level of "fraudulent" emails and affect the security of emails. Therefore, this paper proposes a method based on Bayesian optimization to filter the mail transmission under the hop topology, divide the virtual area and analyze the security level of mail transmission, in order to improve the filtering level of phishing mail in information.

## II. PHISHING MAIL FILTERING THEORY IN INFORMATION SYSTEM

Phishing mail filtering is a program component or subroutine developed by the information system to meet the complex needs of multiple users. Its purpose is to ensure the security of user information and property [5]. Under the hoop topology, the server datacenter of the information system is "virtual partitioned", and $DiskTracker_i$, $ID_j$ and $IP_k$ are set in the partition. Then, a hash table is built in the client node $_l$ of each mail to form a small-scale data search route. Finally, any client node $_l$ performs data filtering by calculating the hash of adjacent clients ($node_{l+1}$, $node_{l-1}$), analyzes "suspicious mail", eliminates "phishing mail", and returns the legal mail *keyValue* in the client. At the same time, each mail is associated with the "virtual partition" to record the address information of legal mail.

### *A.basic Bayesian theory*

Bayesian theory was based on multivariate difference equation, and its mathematical expression is expressed as fellows(1).

$$\begin{cases} node_l = A\square DiskTracker_i + B\square ID_j + \\ keyvalue_j = F\square IP_k + p\sin\left(2k\pi(f_e + \frac{f_D}{f_s})\right)\times T(k) \end{cases} \quad (1)$$

A, B, C and D are coefficient values of corresponding parameters in the information system; $f_e$, $f_s$ and $f_D$ are emails, past emails and feedback results received in the information system datacenter; $p$ is the risk amplification factor of e-mail, which aims to improve the recognition rate of "phishing e-mail", and $T(k)$ or datacenter is the data

actually received by e-mail. Judge whether the mail transmission is safe according to the results of $f_e$, $f_s$ and $f_D$. If $f_e = f_s = f_D \neq 0$ indicates that the phase diagram in the transmission mail will converge at the extreme point, the mail transmission of the information system is safe, $f_e \neq f_s = 0$ , $f_s \neq f_D = 0$ indicates that the transmitted data will converge near the extreme value, and the information system will be attacked or "receive phishing mail".

*B.the Optimization of phishing mail filtering process by Bayesian algorithm*

(1) Receiving of "phishing mail" by the client

The receiving process of "phishing mail" can be described by nonlinear Lagrange equation [7], and the mathematical formula is expressed as fellows(2).

$$\frac{dDiskTracker_i}{dnode_l} = -v'(DiskTracker_i) + F_0 \quad (2)$$

where $g(node_l)$ is the distance between the email and the actually sent data, that is, the possibility that the email received by the client is "phishing email", expressed by Fourier series; $F_0$ represents the total amount of mail received; $v'(DiskTracker_i)$ is a nonlinear discrete function representing mail security.

(2) Server Analysis of "phishing mail"

"Phishing mail" is a key factor in filtering and determines the security degree of the information system [8]. Its mathematical formula is expressed as fellows(3).

$$E[g(node_l)g(node_l + \alpha)] = 6Kg(node_l - \alpha) + \xi \quad (3)$$

where $K$ is the frequency of "phishing mail"; $\xi$ is the error value between receiving mail and sending mail; $E$ is the adjustment coefficient of e-mail security, that is, the coefficient to avoid dangerous e-mail such as "Trojan horse" and "fishing".

*C.Effective mail transfers*

The effective mail transmission amount $F_0$ determines the result of information phishing mail filtering. It is the "safe" mail transmitted from the client to the server. Its mathematical calculation formula is expressed as fellows(4).

$$F_0 = A_0 \cos(2\pi f_0 node_l) \quad (4)$$

where $A_0$ is the function amplitude of data, the parameter $f_0$ is the mail frequency. Both of them jointly determine the "safe" mail volume of mail; Security level = actual data received / total data sent * 100%.

*D.Cache and actual storage*

The mail is put into the cache first. After filtering, it is put into the actual storage area [9] and expressed by a function $v'(DiskTracker_i)$ ,which mathematical calculation formula is expressed as fellows(5).

$$v'(DiskTracker_i) = f_g(\square)^2 + f_h(\square)^4 \quad (5)$$

where $m$ and $n$ are the proportion of "cache" and "actual storage" in the virtual space ; $f_g$ and $f_h$ are "cache" and "actual storage" functions, which are internal algorithms in the hoop topology.

*E.Judgment of critical value*

By deriving formula 5, its inflection point can be obtained, that is

$$v''(DiskTracker_i) = 2\frac{m}{m+n} f_g'(DiskTracker_i) + 4 f_h'(DiskTracker_i)^3 .$$

If formula 4 is substituted, it indicates that the "cache" and "actual storage" are in a two-way balance, that is $A_0 < \sqrt{\frac{4m^2+mn}{27n(m+n)}}$ , the storage resource ratio in the hoop topology is the best. For better "phishing" mail analysis, it will be limited to[- $\sqrt{\frac{m^2+mn}{n(m+n)}}$ , $\sqrt{\frac{4m^2+mn}{27n(m+n)}}$ ].

*F.Rejection of abnormal messages*

The amplification factor is added in the information system to identify the abnormal change information in the mail. If any mail data in the message does not jump, it means that the mail does not contain abnormal data, that is, "phishing data". Otherwise, it is abnormal transmission, and the corresponding data is amplified. By amplifying the abnormal data, the variation amplitude jumps out of $A_0$ .

## III.Building an Improved Bayesian Optimization Model for Phishing Email Filtering

When the information system filters phishing mail, the mail management method in the original hoop topology is too simple. The parsing work of "cache" and "actual storage" is placed in the server to increase the computing load of the server and reduce the management level of "phishing" mail [10]. To solve the above problems, this paper introduces Fourier series, metropolis constraint and Bayesian method to optimize the "phishing" mail filtering, and transfer the parsing work of some servers to the mail client. Among them, Fourier series is used to optimize the calls of "cache" and "actual storage" of e-mail to meet the analysis requirements of "phishing" e-mail[11].

Bayesian method is a non-linear list receiving method. Metropolis acceptance criteria first restricts the mail, and then receives the data center after "filtering" the mail, so as to reduce and improve the recognition rate of "phishing" mail[12]. The specific optimization process is as follows:

*A.Metropolis constraint*

In order to reduce the number of mail filtering, metropolis constraints are applied to determine whether the mail data is "phishing" mail[13]. If the result is "0", it means "phishing" mail. Eliminate or enlarge the value, otherwise conduct corresponding calculation[14]. After the data is constrained by metropolis, assign a value to it to

make the constraint result "0" or "1", and complete the standardization process. According to the above analysis, the mathematical conditions of metropolis constraints are expressed as follows(6).

$$M(\square)=\begin{cases} keyvalue_j \neq keyvalue_{j+1} \\ keyvalue_j = keyvalue_{j+1} \end{cases} \quad (6)$$

where $M(\square)$ is the judgment function of *Metropolis* constraint.

*B. The "phishing" message matches the preset value*

The relationship between datacenter and mail in the server is complex and multidimensional[15]. Mail filtering is affected not only by the number of mail, but also by "cache" and "actual storage". Assuming that the match of "cache" is$P_u$ and the match of "actual storage" is $P_v$, formula 5 can be optimized as follows(7).

$$P\begin{cases} \prod[\sum_{i}^{\infty} v'(\square)+\varsigma]+\sum_{1}^{\infty} mf_g+nf_h & P_u \neq P_v \\ \kappa(P_u-P_{\min})/P_{\max}-P_{\min} & P_u < P_v \\ \lambda(P_v-P_{\min})/P_{\max}-P_{\min} & P_v < p_u \end{cases} \quad (7)$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum bearing matching values of the information system, $\lambda$ is the "cache" adjustment coefficient and $\kappa$ is the "actual storage" adjustment coefficient.

*C.Phishing message filtering and total messages*

In order to filter "phishing" e-mail from massive e-mail, Fourier series is used to analyze "phishing" e-mail and adjust the amount of e-mail, the parsing work of some servers is transferred to the e-mail client, and hash table is introduced to record the "phishing" e-mail website. The original formulas 2 and 3 are optimized into the following formulas(8).

$$\begin{cases} R=\int_{i,j}^{n}[-v'(\square)]+\prod[F_0+\sum_{l}^{\infty}\lim g(\square)]+H_i\square f(node_l) & R=1 \\ U=\int_{i,j}^{n}\sum_{l}^{\infty}\lim g(\square)]+H_i\square f(node_l,g(\square)\} & R=0 \end{cases} \quad (8)$$

$R$ is the number of times mail is received, and its value is 0, indicating that it is a "suspicious" mail. Mail can be analyzed within the group by using the list and eliminated. Otherwise, it can be analyzed by using the view in the hoop topology; U is the number of mail received. Compare the preset mail security level with the actual "suspicious" mail. If the ratio is > 9/10, the mail reception will be cut off, otherwise the mail will be received; $f(node_l, g(\cdot)$ is the calculation function of suspicious degree in "phishing" mail analysis; $H_i$ is the adjustment coefficient of suspicious degree.

*D. Building steps of "phishing" email filtering optimization model*

The simulation model is built according to the data description formula under Bayesian Optimization and multi-user management of information system. The specific steps are as follows:

Under Bayesian optimization, the more frequently the messages are received in the group through the hash table, the smaller the load of the information system and the higher the operation efficiency of the system. Therefore, the optimal critical value is required first, and this value is taken as the threshold for the stable operation of the whole system.

Judge the abnormality of data transmission between e-mail and information system. The abnormality of mail transmission in "phishing" mail filtering is an important optimization index. The occurrence of abnormal data is determined by many factors, such as hash table, integration with hoop topology, division of virtual space in information system, field type, byte size, receiving time, suspicious degree, etc.

Calculate the ratio of "suspicious" mail to actually sent mail.The actual mail sending is the result of user addressing and storage, and is affected by channel, coding, information system modeling (topology, collection, mixing), noise and so on. The "suspicious" e-mail in the information system is consistent with the actual amount of information sent by the e-mail, indicating that the e-mail is a safe e-mail. Among them, the calculation of mail transmission volume is divided into "cache" $P_u$ and "actual storage" $P_v$. During the calculation process, check the mail reception times $R$ and mail reception volume $U$ in the system log to determine whether they are consistent with hoop.

## IV.THE CASE ANALYSIS OF BAYESIAN ALGORITHM FOR PHISHING EMAIL FILTERING OPTIMIZATION IN INFORMATION SYSTEM

*A.The Case introduction*

The software system of the information system is Lniux, the storage server is Lenovo x-3000 series, the storage mode is disk, and the storage space is 723T, The terminal uses wireless communication and several mobile modules to transmit data and send mail. The information system is set to accept mail data and historical data. The data interface is filter signal interface (mainly filtering), mail threshold interface and mail metering port. Through the formula 1 ~ 8 analysis, the "actual storage" filtering optimization model is constructed, as shown in Figure1.

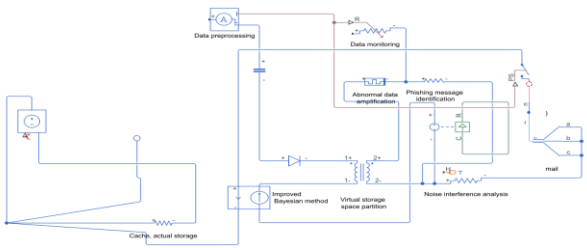


Figure 1. Mail optimization model based on distributed storage method

Take the Hoop topology system of company A as an

example to manage 1,2034 offline emails. After building the mail optimization model of the information system, verify the effectiveness of the model to ensure that it can meet the needs of mail and distributed storage. The results are shown in Figure.2.

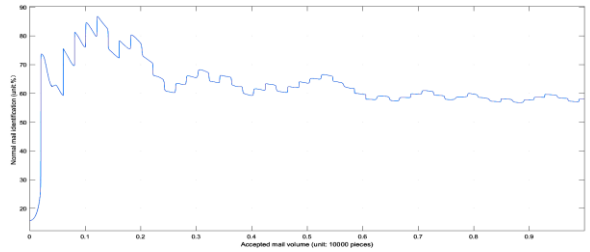

Figure 2. Model verification results

It can be seen from Figure 2 that the mail transmission of the model constructed in this paper is relatively stable, showing regular and equidistant fluctuations, with the maximum amplitude < 7% and the minimum amplitude > 0, indicating that the operation of the model is relatively stable.

*B.the Comparison of optimized results*

The model built by Simulink in MATLAB software is used for analysis, and the results are as follows:

The consistency between receiving "suspicious" e-mail and receiving e-mail in the information system is an important management evaluation index. The results are shown in Figure 3.

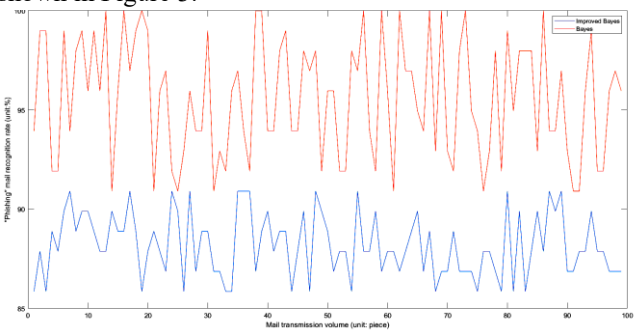

Figure 3. Comparison of compliance between "suspicious" mail and accepted mail

It can be seen from the above figure that the consistency *P* between "suspicious" e-mail and accepted e-mail is good, the consistency rate reached 92.3%. The reason is that before sending data, *Metropolis* analysis is carried out to eliminate "suspicious" mail, and the amplification factor *P* in the hash table is used to identify the suspicious degree.

E-mail security is the focus of phishing e-mail filtering, which involves the privacy of e-mail. Compared with the traditional centralized method, the improved Bayesian method has lower data security. The reason is that there are too many mail terminals and the servers are topologically distributed. Therefore, data security is another indicator to test the management level. The results are shown in Figure 4.

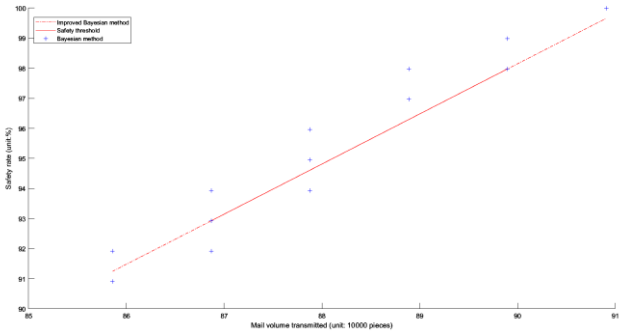

Figure 4. Comparison between improved Bayesian and original Bayesian results

It can be seen from the above figure that the data security level under Bayesian optimization is higher, and the number of secure reception is higher than that of unfiltered method, the average safety rate is 95%. The reason is that under Bayesian optimization, Fourier series is used to limit the number of users on the client, and the extreme value analysis of mail receiving authority is carried out. At the same time, according to the hash table in the mail client, verify the legitimacy of the mail and the rationality of the addressing mode of the receiving end to ensure the security of data transmission.

*C. Virtual area division*

The "Cache" is a significant feature of phishing mail filtering, while "actual storage" is the basic function of the information system, and "cache" / "actual storage" is another indicator of phishing mail filtering. the ratio of "cache" /"actual storage" is higher than that of unfiltered method under Bayesian optimization, which up to 97.2%. This shows that under the same reception times, Bayesian optimization can provide a higher proportion of "cache". As we all know, "caching" is a significant feature of phishing email filtering. Under Bayesian optimization, according to the optimization methods such as hash table and Fourier series, it can improve the processing efficiency of information system for relevant data and reduce the occupation of relevant hardware and software, so as to better provide personalized management services to emails.

## V. Conclusion

Phishing mail filtering is the trend of intelligent development of information system, and Bayesian method can reduce the occurrence rate of "phishing" mail and improve the identification rate of phishing mail in fraud [16]. However, due to the wide application of computers and the increase of the number of mail, the original Bayesian optimization can not meet the requirements [17]. It is a Bayesian Optimization Based on secondary search data table. This method uses data preprocessing and "group" division to improve the processing efficiency of "phishing" mail [18]. In order to make up for the deficiency of Bayesian optimization, Fourier series function and *Metropolis* constraint are added to reduce the occurrence rate of "false eigensolutions"[19].MATLAB simulation results show that the consistency p between the actual mail sending quantity and the received quantity is good, and the data security level, "cache" / "actual storage" ratio and the

number of secure reception under Bayesian optimization are higher than those of unfiltered method. Therefore, the optimization of phishing email filtering in information system is realized under Bayesian optimization, and each important index is better than that under the original Bayesian optimization.

Reference

[1] Cao Lifeng, Lu Xin, Gao Zhensheng, et al. Construction method of e-mail virtual domain isolation based on L - . Journal of communications. 2020,41 (06): 184-201.
[2] Deng Yanli, Jing Hang, Huang can, et al. Research on mail data isolation and encryption. Microcomputer application. 2020,36 (12): 82-85.
[3] Gao Haichao, Chang Yiwen, Yang Wenfeng, et al. Big data operation system based on Hadoop. Science and technology innovation. 2021 (19): 93-94.
[4] Guan Wanqing, Zhang Haijun, Lu Zhaoming. Intelligent resource allocation algorithm for 6G mail network slice based on DRL. Journal of Beijing University of Posts and telecommunications. 2020,43 (06): 132-139.
[5] Huang Danchi, he Zhenwei, Yan Liyun, et al. Research and design of e-mail solution for kubernetes container cloud platform. Telecommunications science. 2020,36 (09): 102-111.
[6] Huang Xiubin, Deng Yanli, Jing Hang, et al. Application Research of phishing mail filtering technology in operation management system. Microcomputer application. 2020,36 (08): 129-131 + 139.
[7] Ji Chengwen, Ma Chao, Zhang Tiegang, et al. Discussion on cloud platform application of Hainan information data center based on microservice and container. Digital communication world. 2020 (12): 172-173 + 184.
[8] Liao Jiawei, sun Yuhua, Wu yonghuan, et al. Research on mail data service platform based on container technology. Information technology and standardization. 2019 (05): 48-52.
[9] Wen Yiyin, Liu Jianxun, Dou Wanchun, et al. Privacy aware email receiving control model in cloud workflow environment. Computer integrated manufacturing system. 2019,25 (04): 894-900.
[10] Xu Li. Simulation of distributed big data cloud storage method based on density evolution. Computer simulation. 2021,38 (07): 424-428.
[11] Zhang Cuicui, sun Jiali, Hong Dehua, et al. Discussion on the application of big data automatic operation and maintenance in power enterprises. Modern industrial economy and informatization. 2020,10 (12): 82-83.
[12] Zhou Zhi, Liu Fangming. Energy efficiency incentive mechanism for resource recycling of mail data center. Chinese Science: information science. 2021,51 (05): 735-749.
[13] Han, N.,D. Liu.PTRE: a probabilistic two-phase replication elimination policy in large-scale distributed storage platforms. Int. J. of Networking and Virtual Organisations.2019, 20(4):23-26.
[14] Luke, L.,A. Jay.The importance of peak pricing in realizing system benefits from distributed storage. Energy Policy. 2021, 15(7):102-104.
[15] R., M., C. N.R., B. J., et al.Distributed storage placement policy for minimizing frequency deviations: A combinatorial optimization approach based on enhanced cross-entropy method. International Journal of Electrical Power and Energy Systems. 2021, 22(3):10-11,.
[16] Yogesh, G. Novel distributed load balancing algorithms in cloud storage. Expert Systems With Applications. 2021, 18(6):23-24.
[17] Shen, T.W., et al. Development and Evaluation of Virtual Reality Induction Electricity Prevention Education and Training Tools for Construction Industry. in 7th IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan). Vol.8.no.7,pp.6-11,2020.
[18] Shi, X.Y., et al., A Didactic Pedagogical Approach toward Sustainable Architectural Education through Robotic Tectonics. Sustainability, vol.12,no.5,pp.7-10,2020.
[19] Torngren, M., F. Asplund, and M. Magnusson, The Role of Competence Networks in the Era of Cyber-Physical Systems - Promoting Knowledge Sharing and Knowledge Exchange. Ieee Design & Test,vol.37,no.6,pp.8-16,2020.

Yahao Zhang, graduated from Beijing University of Technology with a master's degree in computer science in 2018, is currently working in State Grid Cooperation of China Information & Telecommunication Branch as the Intermediate level Penetration Testing Engineer for Headquarters Network and Cloud Environment. main research directions are Cyber Security, Information Security and Artificial Intelligence.

Jin Pang was graduated from North China Electric Power University, Her Major is Computer Science and Technology, She is currently working in State Grid Cooperation of China Information & Telecommunication Branches as the Group Leader for Security Penetration Agency in Security Center. Her main research directions are Industry IoT Security, Penetration Testing and Electric Security.

Hongshan Yin was graduated from Beijing Jiaotong University, Her Major is Information Security, She is currently working in State Grid Cooperation of China Information & Telecommunication Branches as Junior Engineer for Penetration Testing in Electric Control Network. Her main research directions are Industry IoT Security, Penetration Testing and Reverse Engineering .