Phishing Website Detection Algorithm Based on Link Structure

Phishing Website Detection Algorithm Based on Link Structure

Du Shu-Ying[1], He Wang[2, *]

[1]Department of Finance Information, Xuzhou Vocational College of Bioengineering

[2]School of Computer Science and Technology, China University of Mining and Technology

*civihe@foxmail.com

Abstract: Starting from the relationship of link structure, this paper analyses the differences between phishing websites and regular websites. The fingerprints of websites in the collection of suspected target websites are compared and the maximum fingerprint difference is obtained. The threshold of maximum fingerprint difference is set to distinguish phishing websites from regular websites.

1. Introduction

In this paper, the differences between phishing websites and regular websites are analyzed based on the relationship of link structure. Link structure refers to the geometric characteristics between the outgoing and inbound degrees of webpage nodes in Web graphs. Link structure-based phishing website detection algorithm first obtains the suspected target website set according to the suspected target website set [1], then separately analyses the link structure relationship between the suspected target website and the suspected target website set. K-means algorithm is used to quantify the link structure of the website and calculate the link structure fingerprint of the suspected target website set. The fingerprint of the website in the suspected target website set is compared and the maximum fingerprint difference is obtained [2].

2. Phishing Website Detection Algorithm Based on Link Structure

2.1. Suspicious Target Website Set

Attackers make phishing websites look more credible by copying the content of the page of the target website and imitating the appearance of the regular website[3], which not only reduces the cost of making phishing websites, but also makes it easier to deceive users. Therefore, phishing websites and target websites have the same title and keywords. The reason why phishing websites imitate[4] target websites to deceive users is that it is profitable, so the imitated website must be an authoritative website or a central website with high popularity and large number of users. According to the previous chapter, authoritative websites and central websites have high trust value, so the search engine ranks the top in search of related title or keyword target websites.

According to the above analysis, the title and keywords of the website p_{url_0} to be tested are put into the search engine to search[5], and the URLs of the top two sites in the search results are returned as the suspicious target sites set $Turl$=$\{p_{url_1}, p_{url_2}, p_{url_3} p_{url_4}\}$. The title includes the subdomain name of the website URL, the main domain name of the website URL and the content of <title> in the HTML of the website; the keywords refer to CANTINA method[6], and use TF-IDF to get the first three keywords with higher result value as input of the search engine.

2.2. Analysis of Link Structure Characteristics of Web Sites

- Link Interaction Characteristics

In order to make phishing websites more credible, they often link to their target website, while regular websites do not link to phishing websites that imitate them. In web graphs, this feature can be quantified by whether or not the page pointed to by the outgoing link of the web page P under test links back to the web page under test[7]. Link interaction degree f1 is defined to represent this quantization and is calculated by formula (1):

$$f_1 = \sum_{i=1}^{N_{out}} (Nin_{out_i} \times W_{out_i}) \qquad （1）$$

Among them, N_{out} is the number of pages to be tested, out_i is the number of pages to be tested. Formula 2 calculates whether there is such a page in the pages to be measured. In order to avoid the influence of these informal websites, according to whether the outgoing linking website is the suspected target website and whether it is in the normal website collection and promotion website[8] collection, the weights W_{out_i} of the outgoing linking website of the website to be tested are defined. Calculated by formula 3

$$Nin_{out_i} = \begin{cases} 0 & Nin = 1 \\ 1 & Nin = 0 \end{cases} \qquad （2）$$

$Nin = 1$ represents the outbound link of the web page to be tested, and the outbound link of the web page to be tested includes the web page to be tested, while $Nin = 0$ indicates that it does not.

$$W_{out_i} = \begin{cases} 1 & out_i \in Turl \\ Trust(out_i) & out_i \in s^* \cup out_i \in p^- \\ 0 & others \end{cases} \qquad （3）$$

$Trust(out_i)$ is the TrustRank value of out_i.

- Closed Characteristic

Phishing websites use e-mail to transmit false winning information to users. In order to keep the users who enter the website as close as possible, they are often closed[9]. There is a big difference between a closed phishing website and a real website in terms of the number of types of phishing websites. Therefore, the reciprocal of the number of website types (f_2, f_3) is regarded as the closed feature vector.

- Sensitive Page Characteristics

The purpose of phishing website is to collect users' sensitive information and privacy[10], so it must contain pages that require users to submit various account passwords or other personal information. Compared with real websites, the proportion of sensitive pages is less, so the proportion of sensitive pages in the total web pages of websites f_4 is used as feature vector to detect phishing websites.

- Link Structure Quantization

Link structure eigenvector $LStru_{\mathrm{q}}$= (f_1， f_2， f_3， f_4) is defined to represent the link structure feature of webpage q[11]. Then, link structure eigvector $p_{url_i}, p_{url_i} \in \{p_{url_0}, Turl\}, LStru_{p_{url_i^j}}, j \in (0,1,2,\cdots,k-1)$, K is a web site in a Web graph Number of species.$LStru_{p_{url_0^0}}$ is the characteristic vector of the link structure of the website to be tested. And for $LStru_{p_{url_i^j}}, i \in (0,1,2,3,4)$ clustering,$\left\{k_{url_i^n}, i \in (0,1,2,3,4)\right\}$ clustering centers, where n represents the number of cluster centers.

In order to enhance the credibility of phishing websites[13], attackers often link to their counterfeit target websites, some authoritative websites and central websites. In addition, phishing websites often link with other phishing websites to increase the chances of successful phishing. Based on the above analysis, we set the K value to 2.

$$\mathrm{t}_{\mathrm{url}_\mathrm{i}} = \frac{\Sigma_1^4\left(\mathrm{k}_{\mathrm{url}_\mathrm{i}^1} \times \mathrm{k}_{\mathrm{url}_\mathrm{i}^2}\right)}{\sqrt{\Sigma_1^4 \mathrm{k}_{\mathrm{url}_\mathrm{i}^1}{}^2}\sqrt{\Sigma_1^4 \mathrm{k}_{\mathrm{url}_\mathrm{i}^2}{}^2}} \tag{4}$$

Then$T_{url_i} = \left|t_{url_0} - t_{url_i}\right|$, $i \in (1,2,3,4)$, and $\{T_{url_i}\}$. are calculated respectively, and the fingerprint difference set $\{T_{url_i}\}$ of quantifying the link structure relationship is obtained. $MaxT_{url_i}$ is taken as the criterion for judging phishing websites.

3. Parallel Design of K-means Algorithm in Cloud Environment

The main workload of the fishing website detection algorithm based on link structure proposed in this chapter is K-Means clustering process, so K-means clustering algorithm is deployed on cloud platform. Among them, the distance between the data object and the clustering center and the adjustment of the clustering center are two steps that can be calculated in parallel. Therefore, the two steps are processed in parallel with MapReduce framework and serial K-means. Each iteration executes a MapReduce task in parallel K-means.

4. Experiments and Analysis

In order to verify the validity of the detection algorithm of phishing websites based on link structure in this chapter, I set the maximum fingerprint difference between 1000 phishing websites and 2500 ordinary websites as $MaxT_{url_i}$. Then the relationship between the number of phishing websites detected by the two algorithms is analyzed.

4.1. Parameter determination

The maximum fingerprint difference $MaxT_{url_i}$ of 1,000 PhishTank accessible phishing websites and 2,500 regular websites on the whitelist were calculated respectively. By counting the distribution of the maximum fingerprint difference $MaxT_{url_i}$ of each website, it is found that the maximum fingerprint difference distribution of phishing website is shown in Figure 1.

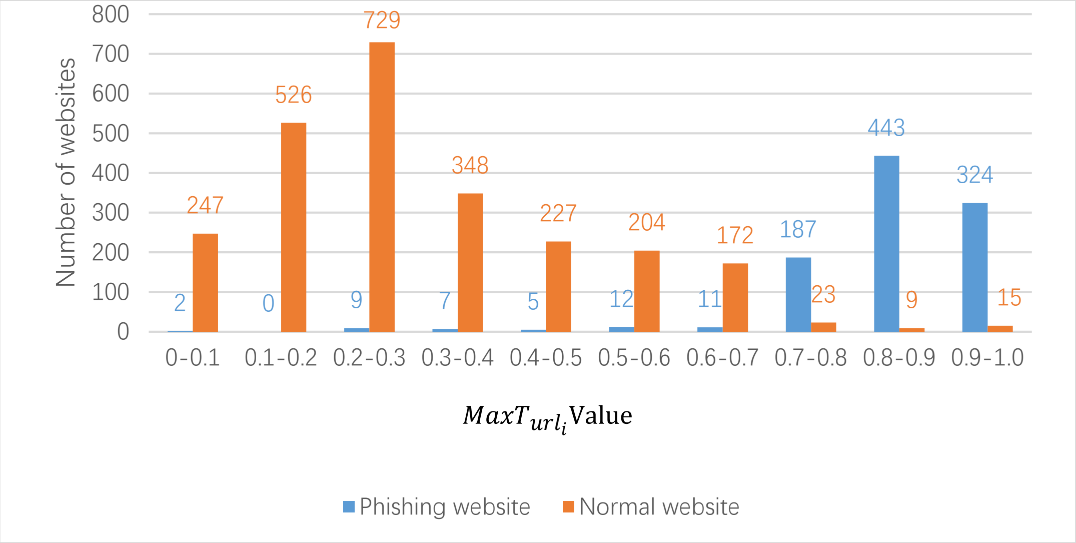

Figure 1 Maximum fingerprint difference

From the data in Figure 1, we can see that the number of phishing websites that can be distinguished from $MaxT_{url_i}$ in (0.7, 1.0) is 954, accounting for 95.4% of the total number of phishing websites, while the number of normal websites that can be distinguished from $MaxT_{url_i}$ in (0, 0.7) is 2453, accounting for 98.12% of the total number of normal websites. Therefore, this paper defines the interval (0.7, 1.0) of $MaxT_{url_i}$ as the threshold interval for detecting phishing websites based on link structure, and judges $MaxT_{url_i}$ as phishing websites in the interval (0.7, 1.0).

4.2. Comparison and Analysis of Experiments

A total of 24,000 phishing websites and 5,000 normal websites were randomly selected. The total number of phishing websites detected by each group quantifies the complementarity rate. The greater the complementarity rate, the more the total number of phishing websites detected by both groups.

$$N_{com} = Num(N_{tl} \cup N_{fl})/Num(p_{fish}) \quad （5）$$

The data in Table 1 are the accuracy rate, misjudgment rate, missed judgment rate and complementary rate of the two algorithms in this experiment.

Table 1 Comparison of experiments

Detection algorithm	Based on improved TrustRank	Link based structure
Accuracy	97.5%	96.4%
False positive rate	0.178%	0.138%
Missing rate	1.78%	2.57%
Complementary rate	98.6%	

In terms of complementarity, because the two methods analyze websites from different angles of links, the types of phishing websites detected by the two methods are quite different, so the total types of phishing websites detected by the two algorithms are larger than that of single phishing websites.

4.3. Stability Analysis of Algorithms

For the experiment of stability analysis, it is also validated in terms of time and quantity. For the stability of time, four groups of experiments were carried out to run the algorithm for 12 hours, 24 hours, 48 hours and 96 hours. The data set of each group of experiments was the data set that the algorithm could detect in 12 hours. Table 2 is the stability data for running different time algorithms.

Table 2 Stability of the algorithm

index time	12h	24h	48h	96h
Accuracy	96.8346%	97.0416%	96.9879%	96.7859%
False positive	0.1346%	0.1358%	0.1337%	0.1356%

```
Function  $U\text{match}(U_p, Uq_i)$
// Input: U_p is the URL of the web page p to be detected, and the URL of the web page in the Uq_i regular website collection.
Input:  $U_p, Uq_i$
// The output is the matching degree of U_p and Uq_i
Output:  $D_{\text{umatch}}$
Begin:
    Count=0;
    $L_{q_i}$=$Uq_i$.length;    //length of $Uq_i$
    Num= Sum($Uq_i$)    // Count the number of Uq_i
    String Maxstr;
    For    i:=1 to Num step 1
        {
            Maxstr=Maxtch($U_p, Uq_i, 2$)  // Find the U_p and Uq_i common subsequence
            Master, the shortest subsequence is 2 in length and stored in Maxstr;
            Lmax= Maxtch.length; // Find the amount of data stored by Maxtch
            For j:=0 to Maxtch.length-1 step 1
                {
                    String str = Maxtch[j];
                    Count+= str.length;
                    D[i-1]=count/($L_{q_i} \times L$max);
                }
            $D_{\text{umatch}}$=Smax(D[ ]) // Take the largest number in the array D
        }
    Print  $D_{\text{umatch}}$
End
```

Figure 2 Algorithm performance

The difference is in the order of thousands of quantities, so is the relationship between the accuracy rate and the trend line of the accuracy rate. Therefore, it can be analyzed that the stability of the algorithm running in the cloud environment in time is better.

For data stability, four groups of experiments are set up. The strategy statistics are used to obtain Table 3:

Table 3 Stability of the algorithm

index test	First set of experiments	Second set of experiments	Third set of experiments	Fourth set of experiments
False positive rate	0.1326%	0.1331%	0.1335%	0.1325%
Accuracy	96.7542%	96.7543%	96.7683%	96.8012%

Although the difference between the third group of experiments is great, the fluctuation range is still at the microdecimal level. The accuracy of each group of experiments differs little from that of the baseline of accuracy, so the algorithm has satisfactory stability in the cloud environment in terms of data quantity.

5. Conclusions of This Chapter

In this chapter, starting from the link structure relationship, by analyzing the link structure relationship between phishing websites and regular websites, we define the link interaction, closeness and sensitive page features combined with page features, and extract the corresponding phishing features of websites to quantify the link structure relationship. We use K-means clustering algorithm to get the link structure fingerprint, and compare the centralized links of suspected target websites. Structural fingerprints are used to determine whether a phishing website is available. Experiments verify the performance and independence of the algorithm in this chapter, and also verify that the algorithm deployed in the cloud environment has relatively satisfactory stability.

References

[1] DOELITZSCHER F, SULISTIO A, REICH C, et al. Private cloud for collaboration and e-Learning services: from IaaS to SaaS[J]. Computing, 2011, 91(1): 23-42.
[2] SHEPHERD D. Containers as a Service (CaaS) is the cloud operatingsystem-i build the cloud. http://www.ibuildthecloud.com/blog/2014/08/19/containers-as-a-servic e-caas-is-the-cloud-operating-system/.
[3] PAHL C. Containerization and the PaaS Cloud[J]. IEEE Cloud Computing, 2015, 2(3): 24-31.
[4] HE H, SHEN H. Green-Aware Online Resource Allocation for Geo-Distributed Cloud Data Centers on Multi-Source Energy[C].Proceeding of the 2017 International Conference on Parallel and Distributed Computing, Applications and Technologies. Piscataway, NJ: IEEE, 2017: 113-118.
[5] ZHAO Y, HUANG Z, and LIU W, et al. A combinatorial double auction based resource allocation mechanism with multiple rounds for geo-distributed data centers[C]. Proceeding of the 2016 IEEE International Conference on Communications. Piscataway, NJ: IEEE, 2016: 1-6.
[6] HUU T T ,THAM C K. An Auction-Based Resource Allocation Model for Green Cloud Computing[C]. Proceeding of the 2013 IEEE International Conference on Cloud Engineering. Piscataway, NJ: IEEE,2013: 269-278.
[7] PRASAD G V, PRASAD A S, RAO S A. Combinatorial Auction Mechanism for Multiple Resource Procurement in Cloud Computing[C]. Proceeding of 2013 International Conference on Intelligent Systems Design and Applications,Piscataway,NJ:IEEE,2013: 337-344.
[8] SHIRLEY S R,P.KARTHIKEYAN. A survey on auction-based resource allocation in cloud environment[J]. International Journal of Research in Computer Applications and Robotics, 2013, 1(9): 96-102.
[9] CHOI Y and LIM Y. Optimization Approach for Resource Allocation on Cloud Computing for IoT[J]. International Journal of Distributed Sensor Networks, 2016, 12(3): 1-6.
[10] ZHAO Y, HUANG Z, and LIU W, et al. A combinatorial double auction based resource allocation mechanism with multiple rounds for geo-distributed data centers[C]. Proceeding of the 2016 IEEE International Conference on Communications, Piscataway, NJ: IEEE,2016: 1-6.
[11] WANG X, CHEN M, LEFURGY C, et al. SHIP: scalable hierarchical power control for large-scale data centers. Proceedings of the 2009 International Conference on Parallel Architectures and Compilation Techniques, 12-16 September 2009, Piscataway, NJ: IEEE, 2009:91-100.
[12] CHOI Y and LIM Y. Optimization Approach for Resource Allocation on Cloud Computing for IoT[J]. International Journal of Distributed Sensor Networks, 2016, 12(3): 1-6.