

Research on Malicious Links Detection System Based On Script Text Analysis

JING Tao*, LI Jun*, XING Rong*

*Network Center of China Science & Technology Network

Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

jingtiao@cstnet.cn

Abstract — Massive trojan and virus spreading via web vulnerabilities has become a severe threat to today's cyber-society. To traditional honeypots, which is able to collect malware with automated manner, becomes one of the most severe and common detection methods nowadays. Compared with honeypots, malicious links detection system based on script text analysis has several advantages due to its lower cost and easier deployment. Firstly, according to the current analysis of trojan web links and existing problems, we summarize the requirement for malicious links detection system, and describe the system design based on requirement. After that we describe the algorithm with mathematical meathod. Finally, we analyze and summarize the experimental results, and verify the reliability and rationality of malicious links by this detection system.

Key words — website security; script text analysis; malicious links

I INTRODUCTION

Nowadays, website security problems have become a serious growing concern, viruses and Trojan spread faster through the sites. Since the public has increased awareness of software security, it is more and more difficult for viruses and Trojan to spread through network software. Now the way that is used by vulnerabilities of host application to execute malicious code has become the main way to spread viruses and malware [1,2]. In this way, hackers use free web space or website intrusion to embed viruses code.

To deal with these security threats, research work for website security technology is constantly evolving. Dr ZHUGE Jian-wei from Institute of Computer Science and Technology, Peking University put an automated malware collection tool based on the high-interaction honeypot principle forward. He focused on the use of honeypot technology to build malicious code automatically capture and collection system [3].

In addition, with recent rapid development of research in webpage text mining [4,5,7], the technology of Trojan detection based on web script code text analysis emerged as the times require. This technology commonly used the method of text feature matching with Trojan detecting rules. Compared with the technology of Trojan detection with honeypot, the technology of website script text feature matching with Trojan detecting rules is suitable for scanning

larger scales of websites due to its lower and shorter time cost [5]. However, this method based on website script text feature relies on the characteristics of the malicious links library in advance. Therefore, this method is difficult to find new malicious web links. And the efficiency and accuracy for the technology of website script text feature with Trojan detecting charactors can not be guaranteed.

In order to predicate the probability for some webpage links are malicious link, we presents an analysis algorithm based on web script text for website security, which can determine probability of webpage malicious links with computing correlation in sets of words from some special webpage scripts code.

With this meathod, we developed a malicious links detection system, which has been verified by experiments that the system can improve the efficiency and accuracy for malicious links detection.

The remnant of this paper is organized as follows: We describe the requirement of this malicious links detection system in section 2. Then we propose the general design of this system. In section 4, we elaborate the mathematical representation of security relevance algorithms and system implementation. Finally, we show the results of the experiments based on this system, and summarize our results and future works.

II SYSTEM REQUIREMENT

The requirement of system is refered to the prevailing scenario and necessity pointed out in the introduction. The basic requirement can be summarized as below:

- System must be able to detect malicious links which have been known.
- System must be able to detect malicious links which have obvious malicious characteristics.
- System must be able to report links which have high probability of malicious links with manual confirmation.
- The system must have a model which is used to calculate the probability of web links being malicious links.

- The model of system which is able to calculate probability of malicious website links must be mathematically feasible.
- System must be able to complete the detection task within an acceptable time.
- System must be able to provide an interface for user query and statistical detection results.

In order to meet the requirement, the system has been designed as discussed in the next section.

III SYSTEM DESIGN

The system is developed to detect malicious links in webpage script and analyze the probability of web links being malicious links. From this point of view, the system is devised as follows:

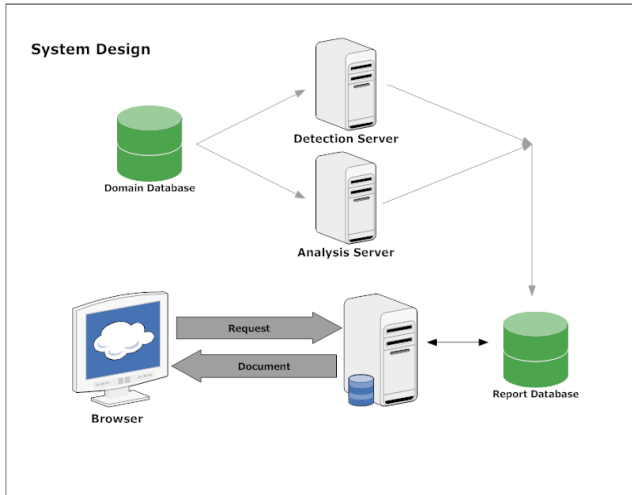


Figure 1. Unsafe links in the webpage text

The overall system architecture is shown in Figure 1. The basic components of the system are:

- Domain database.
Domain database provides domain information of websites that need to be detect if have malicious links.
- Detection Server.
Detection Server detects malicious links in webpage script based on malicious characteristics and library of some known Trojan.
- Analysis Server.
Analysis Server determine the probability of web links being malicious links based on analysis mathematical model, and report links which have high probability of malicious website links .
- Report Database.
Report Database stores reports of malicious links and suspicious malicious links which users are able to query through the interface.

The main innovation of malicious links detection system is analysis algorithm which is used to calculate the probability of malicious website links. So in the next section, we focus on the mathematical model of suspicious malicious links analysis.

Malicious links which hackers implanted are generally to spread malware and Trojan, so degree of the correlation

between website links and contents is low. On the other hand, website links content are appropriate. In view of above analysis, we can obtain the probability of webpage malicious links by analyzing the correlation between website links and contents. In the next section, we will describe the mathematical representation of the specific model which is used to calculate probability of webpage malicious links.

IV Mathematical Representation

A. Related Concepts

1) UL and SL

By analyzing the correlation between web links and web contents[8], the model can be designed by this meathod, we can define some concepts as follows:

Unsafe Links: Web links that have the probability of being malicious links, denoted by

$$UL = \{ul_1, ul_2, ul_3, \dots, ul_m\}$$

Safe Links: Web links that have no probability of being malicious links, denoted by

$$SL = \{sl_1, sl_2, sl_3, \dots, sl_n\}$$

2) Effective Keywords

Unsafe links and safe links cannot be directly used for the calculation of correlation, so define the concept as follows:

Effective Keywords: set of the keywords that can be used to calculate the correlation.

Effective keywords of unsafe links are denoted by

$$KU = \{\{ku_1\}, \{ku_2\}, \{ku_3\}, \dots, \{ku_m\}\}$$

Effective keywords of unsafe links are denoted by

$$KS = \{\{ks_1\}, \{ks_2\}, \{ks_3\}, \dots, \{ks_n\}\}$$

3) AI and NAI

Calculating correlation needs information of web contents, so define the concept as follows:

Authoritative Information : Key information that represents the webpage significance, denoted by

$$KAI = \{kai_1, kai_2, kai_3, \dots, kai_k\}$$

Non-authoritative Information: Information that can't represent the web significance, but is correlated with web contents. In this model, safe links are the non-authoritative information.

B. Calculation of Correlation Modulus

1) Calculation of CMAI

Define the relationship between KU to KAI

$$\mu = \{< \{ku\}, kai > | kai \notin \{ku\}\}$$

Expressed in matrix form:

$$M_\mu = \begin{pmatrix} b_{11} & \dots & b_{1k} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mk} \end{pmatrix} \quad (1)$$

Matrix elements value 1 or 0.

Sum of each row is:

$$M_t = M_\mu * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \quad (2)$$

Define $\{t_1, t_2, t_3, \dots, t_m\}$ as CMAI of unsafe links $\{ul_1, ul_2, ul_3, \dots, ul_m\}$.

2) Calculation of CMNAI

Define the relationship between KU to KS

$$\theta = \{ \langle \{ku\}, \{ks\} \rangle \mid \{ku\} \cap \{ks\} \neq \varnothing \}$$

Expressed in matrix form:

$$M_\theta = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \quad (3)$$

Matrix elements value 1 or 0.

Sum of each row is:

$$M_l = M_\theta * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_n = \begin{pmatrix} l_1 \\ \vdots \\ l_m \end{pmatrix} \quad (4)$$

Define $\{l_1, l_2, l_3, \dots, l_m\}$ as CMNAI of unsafe links $\{ul_1, ul_2, ul_3, \dots, ul_m\}$.

C. Calculation of Correlation

CMW is calculated according to CMAI and CMNAI, which the probability of web links being malicious links is determined by.

1) Calculation of CMW

As in the determination of correlation, authoritative information is more convincing, the weight of CMAI is greater than the weight of CMNAI. So CMW is calculated as:

$$CMW = l_i + mt_i (0 \leq l_i \leq m; 0 \leq t_i \leq k; i = 1, 2, \dots, m) \quad (5)$$

2) Calculation of Correlation

Correlation between unsafe links and web contents is expressed as followed:

$$Correlation = \frac{Correlation \ modulus \ weighted}{Sum \ of \ all \ elements}$$

Then we can draw a formula as below:

$$f(i) = \frac{l_i + mt_i}{m+n+k} (0 \leq l_i \leq m, 0 \leq t_i \leq k; i = 1, 2, \dots, m) \quad (6)$$

$m+n$ is the number of all web links.

Define $f(1), f(2), f(3), \dots, f(m)$ as correlation between unsafe links $\{ul_1, ul_2, ul_3, \dots, ul_m\}$ and web contents.

If the value of correlation is zero, it is indicated that the probability for malicious links is great.

V Schematic diagram of the model

As all above, this section describes schematic diagram of the model which is used to report links which have high

probability of malicious links with manual confirmation. The model is structured into four parts which are shown in Figure II.

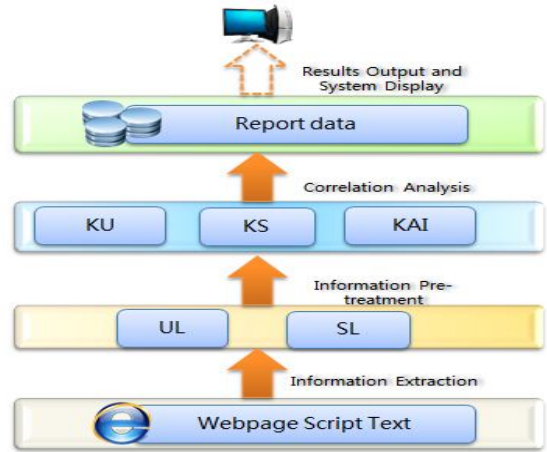


Figure II. Schematic diagram of the model

VI Experimental Analysis

Experimental environment: CPU Intel Core 2 Duo 2.8GHz, Memory 1G, and Windows XP, and as TABLE I.

TABLE I. PARAMETERS OF EXPERIMENTAL ENVIRONMENT

Experimental environment	Parameters
CPU	Intel(R) Xeon(R) CPU E5620 @ 2.40GHz
Memory	48G
Hard Drive	3T
Network Interface Card	Intel PRO/1000 VT Quad Port Server Adapter
Operating System	Red Hat Enterprise Linux Server 5.0

We use website "www.mzfz.gov.cn" as an example. Based on the discussion as above, experiment can be divided into the following specific operations.

A. Information Extraction

Extract the unsafe links, non-authoritative information and authoritative information. We select some of the set of unsafe links:

$$UL = \{qinghe.mzfz.gov.cn, www.454f.com, www.97qq.com, \dots\}$$

We select some of the set of safe links:

$$SL = \{xinban.mzfz.gov.cn / support / sitemap.asp, www.tjrd.gov.cn, changan.mzfz.gov.cn, \dots\}$$

Extract the authoritative information from the website domain (www.mzfz.gov.cn).

B. Information Pre-treatment

Remove the interference of the extracted information and extract sets of effective keywords which can be used to calculate correlation.

Effective keywords of unsafe links are as followed:

$$KU = \{ \{qinghe, mzfz\}, \{454f\}, \{97qq\}, \dots \}$$

Effective keywords of safe links are as followed:

$$KS = \{ \{xinban, mzfz\}, \{jrd\}, \{changan, mzfz\}, \dots \}$$

Effective keywords of authoritative information are as followed:

$KAI=\{mzfz\}$

C. Calculation of Correlation Modulus

According to (1) and (2), calculate CMAI with effective keywords of unsafe links and authoritative information. According to (3) and (4), calculate CMNAI with effective keywords of unsafe links and non-authoritative information.

D. Correlation Analysis

According to (5) and (6), calculate CMW with CMAI and CMNAI, according to which we calculate correlation between unsafe links and web contents.

We select some of the correlation between unsafe links and web contents shown in Table I:

Table II. Correlation between unsafe links and web contents

UL	KU	<i>l</i>	<i>t</i>	<i>f</i>
qinghe.mzfz.gov.cn	{qinghe,mzfz}	210	1	0.83
www.454f.com	{454f}	0	0	0
www.97qq.com	{97qq}	0	0	0
⋮	⋮	⋮	⋮	⋮

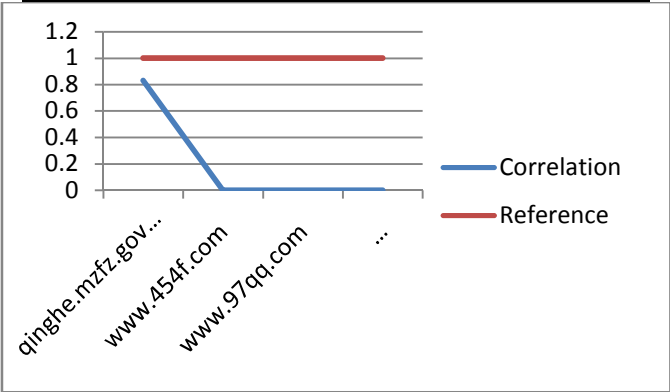


Figure III. Correlation between unsafe links and web contents

From Table II., we can conclude that the correlation of unsafe links(www.454f.com and www.97qq.com) and webpage content is 0, so we determined their probability of being malicious links are high, which in the website script code text are shown in Figure IV.

```
</div><marquee style="width:expression(1);height:expression(1);">
<a href="http://www.yao45.cn">传奇外传私服</a></marquee><div
style="display:none"><a href="http://www.454f.com">传奇私服一条龙</a><a href="http://www.454f.com">私服一条龙</a></div><div
style="position:absolute; left:-999px; top:-999px;"><a title="传
世服" href="http://www.heesf.com/" target="_blank">传世私服</a><a
title="新开传奇世界私服" href="http://chuanqishijie.fsifu.com/"
target="_blank">新开传奇世界私服</a></div><div
style="position:absolute;left:-999px;top:-999px;"> <a href="http:
www.44922.cn/" target=_blank>魔兽私服</a> <a href="http://
www.97qq.com/" target=_blank>魔域SF</a><a href="http://
www.0633dy.com/" target=_blank>魔域私服</a> <a href="http://
www.91219.com/" target=_blank>完美国际私服</a> <a href="http://
www.ilovenmt.cn/" target=_blank>神泣私服</a><a href="http://
www.71530.com/" target=_blank>征途私服</a><a href="http://
www.800qn.cn/" target=_blank>露娜私服</a><a href="http://
www.yztu.cn/" target=_blank>劲舞团私服</a><a href="http://
www.scnw.com.cn/" target=_blank>传奇私服</a><a href="http://
www.5566game.com/" target=_blank>风云私服</a> <a href="http://
www.112399.cn/" target=_blank>魔兽世界私服</a></div><div
```

Figure IV. Unsafe links in the webpage text

These urls are all related with online games or Internet Ads.. They were embedded into webpage by the hackers. These links are malicious. So the algorithm in this paper

can improve the efficiency for the detection of malicious links.

VII Conclusion

In this paper, we propose a malicious links detection system based on webpage script text for website security. And we focus on the model which is used to calculate probability of webpage malicious links. Experimental results showed that the correlation calculated by the model reflects the relationship between web links and webpage contents. Since the model does not consider the text feature of web links, it can improve the efficiency and accuracy of malicious links detection.

In addition, there is a lot of information available for text mining in future. We will work further to optimize the model, including drawing IP addresses and Chinese information factor into text features of web links to improve the efficiency and accuracy of malicious links detection and reduce the false positive rate.

VIII Acknowledgment

This work is supported by Knowledge Innovation Program of Chinese Academy of Sciences, Youth Foundation of Computer Network Information Center, Chinese Academy of Sciences. We should like to thank the help of CNIC for this research.

REFERENCES

[1] Zhuge Jianwei, Holz Thorsten, Han Xinhui, Song Chengyu, Zou Wei. Collecting autonomous spreading malware using high-interaction honeypots. In Lect. Notes Comput. SCI., Zhengzhou, China, December 2007.

[2] LUO Chuan , XIN Mingting , LING Zhixiang. Analysis and realization of the web Trojan horse [J]. NETWORK & COMPUTER SECURITY , Vol.12, pp.83-85, 2007.

[3] Chen Kevin Zhijie, Gu Guofei, Zhuge Jianwei, Nazario Jose, Han Xinhui. WebPatrol: Automated collection and replay of web-based malware scenarios. In Proc. Int. Symp. Inf., Comput. Commun. Secur., ASIACCS, Hong Kong, China, March 2011.

[4] HAN Jia-Wei, MENG Xiao-Feng, WANG Jing, and LI Sheng-En. Research On Web Mining: A Survey [J]. JOURNAL OF COMPUTER RESEARCH & DEVELOPMENT, Vol. 38, No. 4, pp. 405-414, 2001.

[5] E. Glover, K. Tsioutsoulouklis, S. Lawrence, D. Pennock, and G. Flake. Using web structure for classifying and describing web pages. In Proc. of WWW, Hawaii, USA, May 2002. ACM Press.

[6] WU Runpu, FANG Yong, WU Shaohua. Web Trojan Detection Model Based on Statistics and Code Characteristics Analysis [J]. Information and Electronic Engineering , Vol.1, pp.71-75, 2009.

[7] A. Sun and E.-P. Lim. Web unit mining – finding and classifying subgraphs of web pages. In Proc. of 12th ACM CIKM, pages 108–115, New Orleans, LA, USA, Nov. 2003.

[8] Kleinberg J M. Authoritative sources in a hyperlinked environment .Journal of ACM, 1999, 46(5) :604~632