*I.J. Computer Network and Information Security*, **2011, 3, 41-47**
Published Online April 2011 in MECS (http://www.mecs-press.org/)

# Design and Implementation for Malicious Links Detection System Based On Security Relevance of Webpage Script Text

XING Rong
Network Center of China Science & Technology Network
Computer Network Information Center, Chinese Academy of Sciences
Beijing 100190, China
Email: xingrong@cstnet.cn

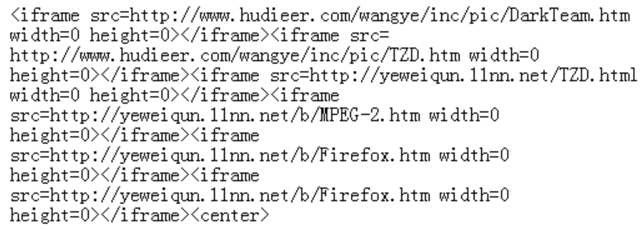
LI Jun and JING Tao
Network Center of China Science & Technology Network
Computer Network Information Center, Chinese Academy of Sciences
Beijing 100190, China

***Abstract* —With the development of web technology, spreading of Trojan and viruses via website vulnerabilities is becoming increasingly common. To solve this problem, we propose a system for malicious links detection based on security relevance of webpage script text and present the design and implementation of this system. Firstly, according to the current analysis of malicious links, we describe requirements and the general design for detection system. Secondly we describe the security-related algorithm with mathematical language, and give the data structure of this algorithm. Finally, we analyze and summarize the experimental results, and verify the reliability and rationality of system.**

***Index Terms*—text analysis; security relevance; malicious links**

## I INTRODUCTION

With the popularity of web technology, website security problems become more and more popular, viruses and Trojan spread faster through the sites. The way that viruses and Trojan spread through the network software has becomes difficult. Now using vulnerabilities of host application to execute malicious code has become the main way to spread viruses and malware [1, 2]. Hackers use free web space or website intrusion to embed virus's code. The webpage script code text of web Trojan links is shown as Fig. 1:

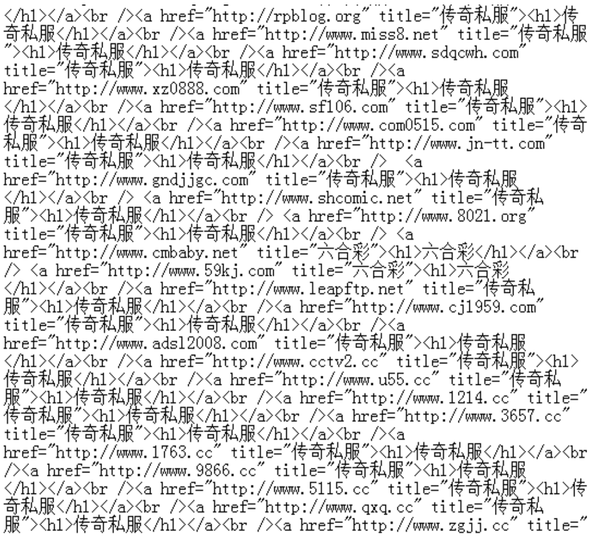


```
<iframe src=http://www.hudieer.com/wangye/inc/pic/DarkTeam.htm
width=0 height=0></iframe><iframe src=
http://www.hudieer.com/wangye/inc/pic/TZD.htm width=0
height=0></iframe><iframe src=http://yeweiqun.11nn.net/TZD.html
width=0 height=0></iframe><iframe
src=http://yeweiqun.11nn.net/b/MPEG-2.htm width=0
height=0></iframe><iframe
src=http://yeweiqun.11nn.net/b/Firefox.htm width=0
height=0></iframe><iframe
src=http://yeweiqun.11nn.net/b/Firefox.htm width=0
height=0></iframe><center>
```

Figure 1.Web Trojan links in the webpage script text

Some links in the webpage are not web Trojan links, but they are malicious links which Hackers embed into the webpage script text to improve their own sites ranking. These malicious links also pose a threat to website security and they are usually related with online games or Internet advertisement. The webpage script code text of malicious links is shown as Fig.2:

*This work is supported by Knowledge Innovation Program of Chinese Academy of Sciences, Youth Foundation of Computer Network Information Center, Chinese Academy of Sciences.(Grant No. CNIC_QN_10001)



```
</h1></a><br /><a href="http://rpblog.org" title="传奇私服"><h1>传
奇私服</h1></a><br /><a href="http://www.miss8.net" title="传奇私服
"><h1>传奇私服</h1></a><br /><a href="http://www.sdqcwh.com"
title="传奇私服"><h1>传奇私服</h1></a><br /><a
href="http://www.xz0888.com" title="传奇私服"><h1>传奇私服
</h1></a><br /><a href="http://www.sf106.com" title="传奇私服"><h1>
传奇私服</h1></a><br /><a href="http://www.com0515.com" title="传奇
私服"><h1>传奇私服</h1></a><br /><a href="http://www.jn-tt.com"
title="传奇私服"><h1>传奇私服</h1></a><br />  <a
href="http://www.gndjjgc.com" title="传奇私服"><h1>传奇私服
</h1></a><br /> <a href="http://www.shcomic.net" title="传奇私
服"><h1>传奇私服</h1></a><br /> <a href="http://www.8021.org"
title="传奇私服"><h1>传奇私服</h1></a><br /> <a
href="http://www.cmbaby.net" title="六合彩"><h1>六合彩</h1></a><br
/> <a href="http://www.59kj.com" title="六合彩"><h1>六合彩
</h1></a><br /><a href="http://www.leapftp.net" title="传奇私
服"><h1>传奇私服</h1></a><br /><a href="http://www.cj1959.com"
title="传奇私服"><h1>传奇私服</h1></a><br /><a
href="http://www.ads12008.com" title="传奇私服"><h1>传奇私服
</h1></a><br /><a href="http://www.cctv2.cc" title="传奇私服"><h1>
传奇私服</h1></a><br /><a href="http://www.u55.cc" title="传奇私
服"><h1>传奇私服</h1></a><br /><a href="http://www.1214.cc" title="
传奇私服"><h1>传奇私服</h1></a><br /><a href="http://www.3657.cc"
title="传奇私服"><h1>传奇私服</h1></a><br /><a
href="http://www.1763.cc" title="传奇私服"><h1>传奇私服</h1></a><br
/><a href="http://www.9866.cc" title="传奇私服"><h1>传奇私服
</h1></a><br /><a href="http://www.5115.cc" title="传奇私服"><h1>传
奇私服</h1></a><br /><a href="http://www.qxq.cc" title="传奇私
服"><h1>传奇私服</h1></a><br /><a href="http://www.zgjj.cc" title="
```

Figure 2.Malicious links in the webpage script text

To deal with these security threats, research work for website security technology is constantly evolving. Many studies focused on Honeypot technology to address security threats [11, 12, 13]. In addition, with the development of research in webpage text mining [4, 6, 7, 8, 9], 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 large scales of websites with its low cost and short time cost [5]. However, this method of website script text feature relies on the characteristics of the malicious links library in advance, so it’s difficult to find the newer malicious web links. The efficiency and accuracy for the technology of website script text feature with Trojan detecting rules can not be guaranteed.

In order to predicate the probability for some webpage links are malicious link, we present an analysis system 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 text, and is verified by experiments that the system can improve the efficiency and accuracy for malicious links detection.

The remainder of this paper is organized as follows: We describe related work in section 2. Then we present requirement of the system in section 3. In section 4, we propose the general design of this system. Then we propose the mathematical representation of security relevance algorithms and design the data structure of algorithms. In section 8, we show the results of the experiments based on this system. Finally, we summarize our results and future works in section 9.

## II RELATED WORK

In general, current studies for malicious links detection are classified into two categories: The first one detects malicious links based on honeypot principle, while the other one uses the characteristics of the malicious links. 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]. CHEN Ling from Institute of Information Security Engineering, Shanghai Jiaotong University presented a Web Trojan detection scheme based on HoneyClient which is an efficient client-side honeypot system[14]. WU Run-pu from Institute of Information Security, Sichuan University proposed a Web Trojan detecting model based on statistics and analysis of code characteristics [10].

Research studies on webpage text mining have recently gained lots of attentions. But, current studies of webpage script text feature matching with Trojan detecting rules focus on the characteristics of malicious links, there are lots of information for Web mining. So we propose a system for malicious links detection based on security relevance of webpage script text and present design and implementation of this system.

## III SYSTEM REQUIREMENT

The requirement of system is based on 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 characteristics.
- System must be able to report links which have high probability of being malicious links.
- The system must have a model which is able to calculate the probability of web links being malicious links.
- The model of system must be feasible in the level of mathematics.
- System must be able to complete the detection task within an acceptable time.
- System must be able to provide a query interface for users.

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

## IV 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 Fig.3:

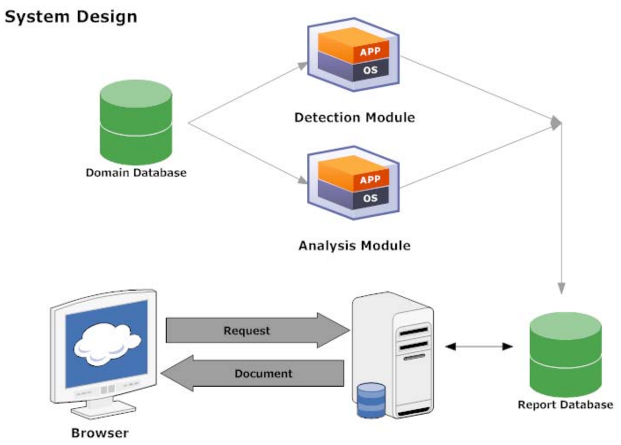


Figure 3.General design of the detection system

The overall system architecture is shown in Fig. 3. The basic components of the system are:

- Domain database.
  Domain database stores webpage script texts of websites that need to be detect if have malicious links.
- Detection Module.
  Detection Module detect malicious links in webpage script based on known characteristics
- Analysis Module.
  Analysis Module determine the probability of web links being malicious links based on analysis mathematical model.
- Report Database.
  Report Database stores reports of malicious links and suspicious malicious links which are used to provide a query interface for users.

In analysis module, we introduce the concept of correlation to analyze malicious links. Correlation determines the possibility of a link being malicious link. So the general flow chart of the system is shown as Fig. 4:

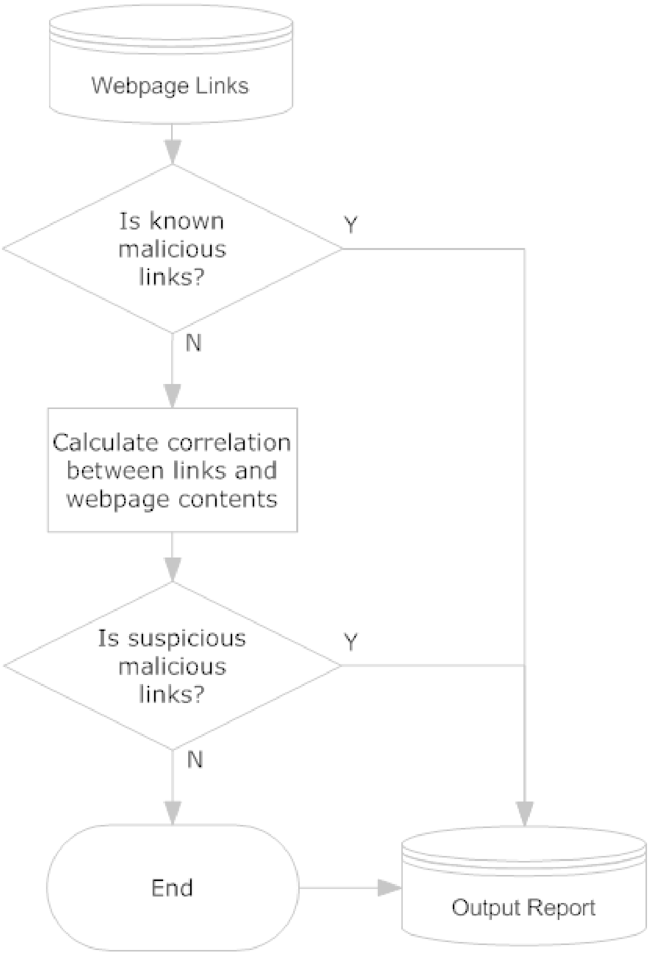

Webpage Links
Is known malicious links?
Y
N
Calculate correlation between links and webpage contents
Is suspicious malicious links?
Y
N
End
Output Report

Figure 4.Flow chart of the detection system

This article focuses on the analysis of suspicious malicious links. So in the next section, we introduce the mathematical model of suspicious malicious links analysis.

## V MATHEMATICAL REPRESENTATION OF ANALYSIS MODULE

### *A. Related Concepts*

*1) UL and SL*

By analysing the correlation between web links and web contents, the model determines probability of webpage malicious links, so we can define the concept as follows:

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

$$UL = \{\mathrm{ul}_1, ul_2, ul_3, \cdots, ul_m\}$$

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

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

*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

$$KU = \{\{\mathrm{ku}_1\},\{ku_2\},\{ku_3\},\cdots,\{ku_m\}\}$$

Effective keywords of unsafe links are denoted by

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

### 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

$$KAI = \{\mathrm{kai}_1,\mathrm{kai}_2,\mathrm{kai}_3,\cdots,\mathrm{kai}_k\}$$

Non-authoritative Information：Information that can't represent the web significance, but it 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

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

Expressed in matrix form:

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

Matrix elements value 1 or 0.

Sum of each row is:

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

Define $\{t_1,t_2,t_3,\cdots,t_m\}$ as CMAI of unsafe links $\{\mathrm{u}l_1,ul_2,ul_3,\cdots,ul_m\}$.

### 2) Calculation of CMNAI

Define the relationship between KU to KS

$$\theta = \{<\{\mathrm{ku}\},\{ks\}> \left|\{ku\}\cap\{ks\} \neq \varphi\right.\}$$

Expressed in matrix form:

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

Matrix elements value 1 or 0.

Sum of each row is:

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

Define $\{l_1,l_2,l_3,\cdots,l_m\}$ as CMNAI of unsafe links $\{\mathrm{u}l_1,ul_2,ul_3,\cdots,ul_m\}$.

## 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:

$$CMW = l_i + mt_i (0 \le l_i \le m; 0 \le t_i \le k; i = 1,2,\cdots,m) \qquad （5）$$

### 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:

$$f(i) = \frac{l_i + mt_i}{m+n+mk} (0 \le l_i \le m; 0 \le t_i \le k;$$
$$i = 1, 2, \mathrm{L}\ \ , m) \qquad （6）$$

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

Define $f(1), f(2), f(3),\cdots, f(m)$ as correlation between unsafe links $\{\mathrm{u}l_1,ul_2,ul_3,\cdots,ul_m\}$ and web contents.

The more correlation close to zero, indicating that the greater probability of web links being Malicious links is, vice-versa.

## VI Model Description Of Analysis Module

Model describes the relationship between the elements (unsafe links, safe links, authoritative information, non-authoritative information and webpage content), the model is defined as a quintuple:

A (Algorithm) = {UL, SL, AI, NAI, WC}

-UL(Unsafe Links)：Web links that have the probability of being determined malicious links

-SL(Safe Links)：Web links that have no probability of being malicious links

-AI(Authoritative Information)：Key information which can represents the webpage significance

-NAI(Non-Authoritative Information)：Information that can't represent the webpage significance, but it is correlated with webpage contents. In this model, safe links are the non-authoritative information.

-WC(Webpage Script Text Content)：Information of the webpage script code text contents

$$UL = \{L_1, L_2, L_3,\cdots, L_{\mathrm{m}}\}$$

$$-L(links) = \{CMAI, CMNAI, CMW\}$$

---CMAI: Correlation modulus for authoritative information

---CMNAI: Correlation modulus for non-authoritative information

---CMW: Correlation modulus for web

## VII DESIGN OF ANALYSIS MODULE

Based on the above discussion, analysis module is designed as followed. Module process is divided into four steps:

Step 1. Information Extraction

Extract the unsafe links, non-authoritative information and authoritative information from the webpage. In this model, safe links are the non-authoritative information.

Step 2. Information Pre-processing

Remove the interference words from the extracted information which can be used to calculate correlation, such as www, cn, us, de, and so on.

For example, if the extracted information is “www.google.com.hk”, the words “www”, “hk” and dot are the interference words.

Step 3. Calculation of Correlation Modulus

Calculate CMAI according to the correlation between unsafe links and authoritative information, and calculate CMNAI according to the correlation between unsafe links and non-authoritative information. Model specifically addressed as follows:

$$\begin{cases} if \;\; L_{\text{i}} \cap AI \neq \phi, \;\; then \;\; L_{\text{i}}.CMAI = Max \\ if \;\; L_{\text{i}} \cap NAI \neq \phi, \;\; then \;\; L_{i}.CMNAI = UL.CMNAI + 1 \end{cases}$$

（$i=1, 2, 3, \cdots, m$）

Flow chart of core code for calculating CMAI is shown in Fig. 5:

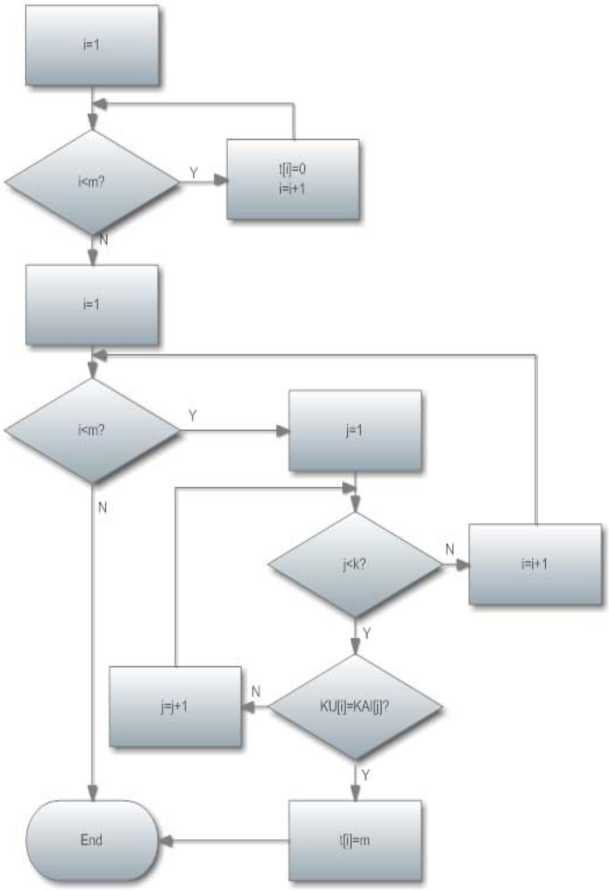


Figure 5.Flow chart of core code for calculating CMAI

Flow chart of core code for calculating CMNAI is shown in Fig. 6:

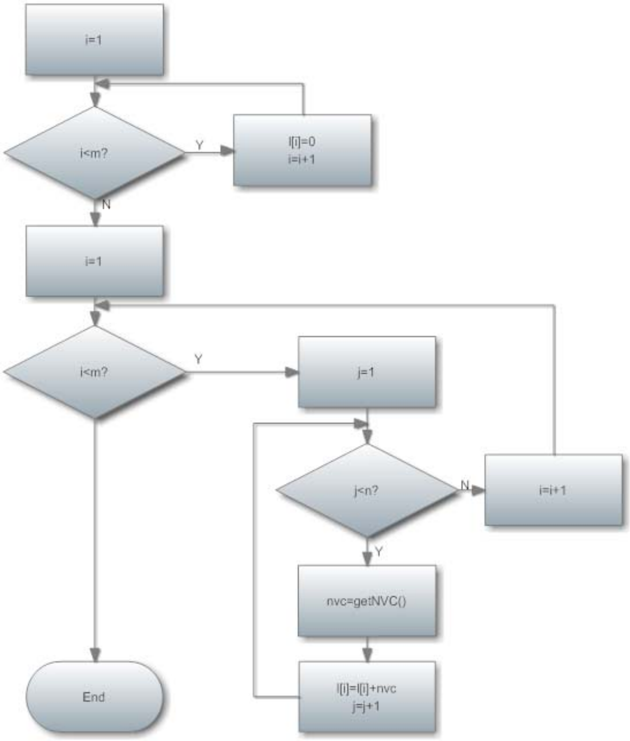


Figure 6. Flow chart of core code for calculating CMNAI

Function getNVC() is to calculate the correlation between one unsafe link and all safe links, of which flow chart is shown in Fig.7:

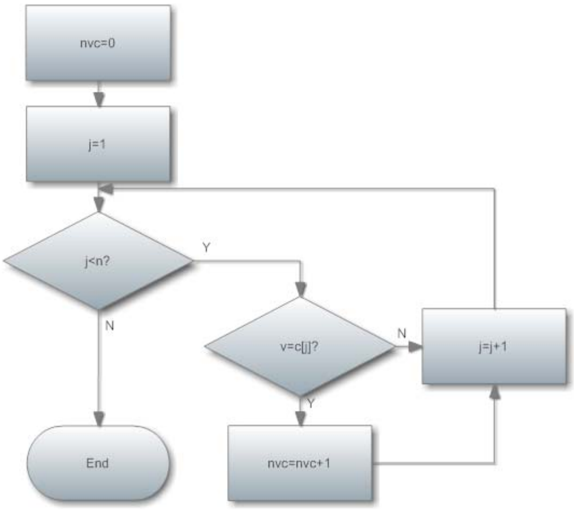


Figure 7. Flow chart of function getNVC()

Step 4. Correlation Analysis

Calculate CMW according to CMAI and CMNAI, which the probability of webpage malicious links is determined by.

## VIII EXPERIMENTAL ANALYSIS

The main parameters of experimental environment are shown 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 a website “colors.dufe.edu.cn” as an example. Based on the above discussion, 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:

$$UL = \{newspaper.dufe.edu.cn, www.80xiazai.com, www.zhaosfpk.com, \mathrm{L}\ \}$$

We select some of the set of safe links:

$$SL = \{www.dufedsn.com/dsn/, showtime.dufe.edu.cn, youth.dlmu.edu.cn, \mathrm{L}\ \}$$

Extract the authoritative information from the website domain (colors.dufe.edu.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:

$$KU = \{\{newspaper, dufe\}, \{80xiazai\}, \{\mathrm{zhaosfpk}\}, \mathrm{L}\ \}$$

Effective keywords of safe links are as followed:

$$KS = \{\{dufedsn\}, \{showtime, dufe\}, \{youth, dlmu\}, \mathrm{L}\ \}$$

Effective keywords of authoritative information are as followed:

$$KAI = \{\mathrm{colors,dufe}\}$$

### C. Calculation of Correlation Modulus

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

### D. Correlation Analysis

According to the formula (5) and formula (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 II and Fig. 8:

TABLE II.CORRELATION BETWEEN UNSAFE LINKS AND WEB CONTENTS

| UL | KU | l | t | f |
|---|---|---|---|---|
| newspaper.dufe.edu.cn | {newspaper,dufe} | 91 | 1 | 0.62 |
| www.80xiazai.com | {80xiazai} | 0 | 0 | 0 |
| www.zhaosfpk.com | {zhaosfpk} | 0 | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

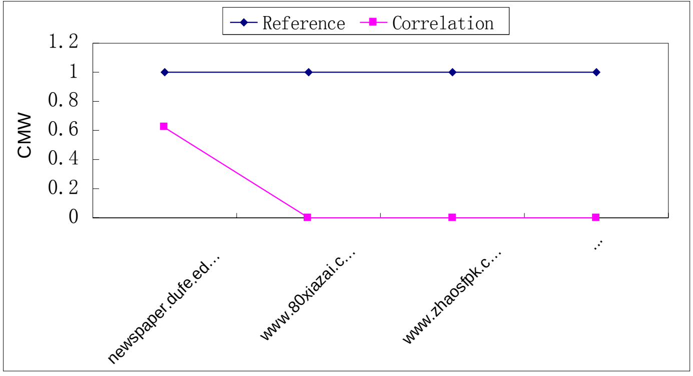
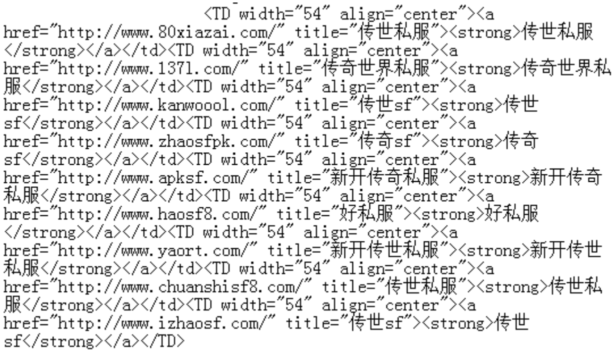


Figure 8. Correlation between unsafe links and web contents

From Table II, we can conclude that the correlation of unsafe links(www.80xiazai.com and www.zhaosfpk.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 Fig. 9.



```
                        <TD width="54" align="center"><a
href="http://www.80xiazai.com/" title="传世私服"><strong>传世私服
</strong></a></td><TD width="54" align="center"><a
href="http://www.1371.com/" title="传奇世界私服"><strong>传奇世界私
服</strong></a></td><TD width="54" align="center"><a
href="http://www.kanwoool.com/" title="传世sf"><strong>传世
sf</strong></a></td><TD width="54" align="center"><a
href="http://www.zhaosfpk.com/" title="传奇sf"><strong>传奇
sf</strong></a></td><TD width="54" align="center"><a
href="http://www.apksf.com/" title="新开传奇私服"><strong>新开传奇
私服</strong></a></td><TD width="54" align="center"><a
href="http://www.haosf8.com/" title="好私服"><strong>好私服
</strong></a></td><TD width="54" align="center"><a
href="http://www.yaort.com/" title="新开传世私服"><strong>新开传世
私服</strong></a></td><TD width="54" align="center"><a
href="http://www.chuanshisf8.com/" title="传世私服"><strong>传世私
服</strong></a></td><TD width="54" align="center"><a
href="http://www.izhaosf.com/" title="传世sf"><strong>传世
sf</strong></a></TD>
```

Figure 9.Unsafe links in the webpage text

These links are all related with online games or Internet advertisement. They were embedded into webpage by the hackers. These links are malicious. So the system in this paper can improve the efficiency for the malicious links detection.

## IX CONCLUSION

In this paper, we propose a system for malicious links detection based on security relevance of webpage script text. 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.

## REFERENCES

[1] ZHUGE Jianwei, YE Zhiyuan, ZOU Wei. “Research on Classification of Attack Technologies” [J]. Computer Engineering，Vol.31, No.21, pp.121-126, 2005.

[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] ZHUGE Jianwei，HAN Xinhui，ZHOU Yonglin. HoneyBow: an automated malware collection tool based on the high-interaction honeypot principle [J]. Journal on Communications，2007,28(12):8-13.

[4] E. Glover, K. Tsioutsiouliklis, 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.

[5] 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.

[6] 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.

[7] Han J, Kamber M. Data Mining: Concepts and Techniques SanMateo, CA: Morgan Kaufmann, 2000.

[8] 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.

[9] XUE Wei-min and LU Yu-chang. Research On Text Data Mining [J]. Journal of Beijing Union University(Natural Sciences), Vol. 19, No. 4, pp. 59-63, 2005.

[10] 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.

[11] LEVINE J, GRIZZARD J, OWEN H. Application of a methodology to characterize rootkits retrieved from honeynets[A]. Proceedings of the Fifth Annual Information Assurance Workshop[C]. West Point, NY, USA, 2004. 15-21.

[12] PROVOS N. A virtual honeypot framework [A]. Proceedings of 13th USENIX Security Symposium[C]. San Diego, CA, USA, 2004. 1-14.

[13] Ali Ikinci， Thorsten Holz， Felix Freiling. Monkey-Spider：Detecting Malicious Websites with Low-Interaction HoneyClients[C]//Gesellschaft für Informatik. Proceedings of Sicherheit. Mannheim：University Mannheim，2008：233-244.

[14] CHEN Ling, WANG Yi-jun and XUE Zhi. Detection of Web-site with Trojan Based on HoneyClient [J]. CHINA INFORMATION SECURITY, Vol. 5, pp. 75-77, 2010.

**XING Rong**, born in 1988, master candidate, his research interests include network security and network management.

**LI Jun**, born in1968, professor, Ph. D. supervisor, his research interests include network security and next generation network.

**JING Tao**, born in 1979, Ph. D. candidate, his research interests include network security and intrusion detection.