PHISHING SITES DETECTION BASED ON URL CORRELATION

Ying Xue[1,2], Yang Li[1,2], Yuangang Yao[2], Xianghui Zhao[2], Jianyi Liu[1], Ru Zhang[1]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]China Information Technology Security Evaluation Center, Beijing, 100085,China
xue14151@163.com, ly9xyzn@foxmail.com, yaoyg@itsec.gov.cn, zhaoxh@itsec.gov.cn, liujy@bupt.edu.cn, zhangru@bupt.edu.cn

Abstract: With the rapid development of the information technology, internet security has drawn more and more attention. Nowadays, most researchers focus on lexical and host features to classify Phishing URLs. In this paper, we proposed Vulnerable Sites List and a new feature which is named URL Correlation. URL Correlation is based on the similarity of URLs with the List above that we created. In addition, a large improvement of accuracy is observed by comparing methods which use our new feature with the others which use the normal one.

Keywords: URL Correlation; Vulnerable Sites List; Weighting; Machine Learning Method

1 Introduction

Nowadays, the number of phishing URLs have grown a lot. The new comers have increased by 5.59 million at the first half of the year 2014 [1]. Meanwhile, phishing sites have attracted people's growing attention.

A lot of good suggestions are proposed by many reseachers. The main methods include High interaction honeypot technology, HTML static checking algorithm and Machine learning technology [2]. Although the static checking algorithm [3] is very fast, it also exists the problem of low accuracy. The accuracy of honeypot technology [4] is high, however, the requirements of the hardware are also high. Machine learning method, with a good accuracy and time complexity, is a good choice.

Most researchers improved the accuracy of classification via feature selection. Ma et al. [5] propose a new classifier, which includes lexical features (like the length, special number) and host-based features (like WHOIS). Chen Zhuang et al. [6] not only use familiar features in lexical features and host-based features, but also propose a new method to make full use of the information of registration authority, through the algorithm of WOE. Joby James et al. [7] introduce a new property which is named the Page Importance.

As we all know, most phishing sites just imitate few popular URLs which people browse and log in. So, in this paper, we take full advantage of that. At first, we search for the top URLs in Google's PageRank [8], and also put the URLs which are closely related to money and privacy, like http://www.psbc.com and the URLs which are usually imitated by the malicious attackers into the Vulnerable Sites List. We define the different distances in URLs based on the distances in strings, which is named Levenshtein and also put forward a new method to calculate the same distances between 2 URLs. Considering the length's influence on the results, we introduced the rate of the same distances and different distances. Combined with the above methods, we present a new criterion which is named URL Correlation. At last, we introduced the method of weighting to find the best matching URLs.

The contributions of this paper are:

◎A Vulnerable Sites List, including domains which are the most frequently targeted or some with a top PageRank is created.

◎A new method to calculate the correlation between domains is introduced.

◎Through putting weights, we turn multi-objective programming to single objective programming to find the best match URLs for any given sites.

The remaining of the paper is structured as follows. In section 2 we introduce the normal feature selection method. In section 3, we describe the feature we put forward. In section 4 we compare the accuracy when we use our feature with those without using this method. At last, in section 5 we have a conclusion.

2 Normal Feature Selection method

In most of the phishing attacks, the evildoers will confuse users through adding some special characteristics or delete some words. And also, for a phishing sites, the registration time is often short and rarely updated.

To analyze the basic features of URLs, most of the researchers choose a combination of lexical and host based features.

Lexical Features

In a phishing attack, the evildoers will cheat victims through obfuscating phishing URLs with the benign one. There are many correlations between the benign URL and the phishing one. The main features are listed as follows, which include the length of the domains, the length of entire URL, as well as the number of dots. In order to confuse users, phishing sites are often used with

978-1-5090-1256-5/16/$31.00 ©2016 IEEE

a lot of dots, like http://ebay.com.register.online-service.bank.login/... And they will also use many special characters like "/","&","~", which can be rarely found in the normal sites. Besides, a considerable number of capital letters can be found in some phishing sites.

Host Based Features

WHOIS properties—which include the date of registration, the date of expiration and the date of update, the registrar and the registrant [5].In order to make full use of the WHOIS information, we introduce three arguments time1, time2, time3, where time1 is the interval between expiration time and present time, time2 is difference of expiration and expiration and time3 is the deviation of update and present time.

3 New Features extraction based on the URL Correlation

Nowadays, most researchers choose to find features from lexical features, PageRank and WHOIS information. Ma et al. [5] use the combination of Lexical Features and Host information. Whittaker et al. [9] use a scalable machine learning classifier based on the Google's phishing blacklist. Blum et al. [10] use the lexical features, hostname biagrams. Feroz [11] uses the URL Ranking Feedback, besides the Lexical Features, Bigrams and Host. While, Garera et al. [12] introduce a new method, the red flag, which includes the words that the phishing sites often imitate. Some use the contend-based approach, which is too complex.

There are many indices for objective evaluation. As we analyze the dataset, and find that, in most times, it is well-organized. It usually deletes some word, add some characters, or some special symbols to confuse users. Usually, phishing URLs are from the benign ones through making some changes to them. For example, the benign one is www.ebay.com, while the phishing one is www.ibay.com, or www.bay.com, or even www.ebay.net. So, the more similar with the URLs (but not completely equal) in the Vulnerable Sites List, is more likely to be the malicious sites. Depending on all about that, we put forward a new criterion, which is named URL Correlation. This criterion includes the rate of different distances and same distances. We use the Levenshtein and combine it with the trait of URL to define the different distance and same distance. The Levenshtein distance is proposed by Vladimir Levenshtein [13], a Russian scientist.

3.1 Vulnerable Sites List construction

Phishing targets are widely found, but the main targets are very concentrated. Only a few well-known sites become the targets of more than half of phishing URLs, Such as www.paypal.com, www.tibia.com [13].

In China, the total number of top 7 vulnerable targets account for 95.1% in the total number of phishing targets which are reported [6]. We choose URLs where their PageRank are in the top of Google's PageRank [8]. And also, we choose the top sites which are usually attacked, imitated by malicious attackers or closely related to money and privacy.

3.2 Definition of Distance

Take a URL as an example, such as www.ebay.com/login... www.ebay.com is defined as the domains, and the next section is the path. In our Model, we introduce a new criterion named URL Correlation which is based on the domains. This criterion includes two indexes, the rate of different distances and same distances. The different distances in 2 URLs are defined as *diff*, the rate is *diff_rate*. The different distance between 2 URLs is the minimum number of single-character edits, which include the number of insertions, deletions or substitutions. We define them as insertions, deletions, and substitutions. The same distance is defined as same, the rate regards as *same_rate*.

Assume the domains is $U=u_1u_2,..., u_m$, aim domain is $V=v_1v_2,..., v_n$, their lengths are n and m. The total length of the Aim String is defined as *len, len=length(V)*. Based on the Levenshtein distance, we define the different distance between 2 URLs. Calculation formula is as follow.

if $U_i=0, V_j=0,$ *then* $Leven\ (U_i, V_j)=0;$

if $U_i=0, V_j>0,$ *then* $Leven\ (U_i, V_j)=V_j;$

if $U_i>0, V_j=0,$ *then* $Leven\ (U_i, V_j)=U_i;$

if $U_i>=1, V_j<=1,$ *then* $Leven(U_i, V_j))=min\{Leven(U_{i-1,} V_j)+1, Leven\ (U_{i,} V_{j-1})+1, Leven(U_{i-1,} V_{j-1})+cov(U_{i,} V_j)\},$

$U=U_1U_2,...,Um; V=V_1V_2,...,Vn;$

$Cov(U_i, V_j)=1,$ *if* $U_i,!=V_{j;}$ *Else,* $Cov(U_i, V_j)=0,$ *if* $U_i,=V_j$

Figure 1 Process to calculate the distances in two domains

Then *diff* and *diff_rate* is defined like this

$$diff=Leven(U,V)$$

$$diff\_rate=diff/len \qquad (1)$$

The same length should be less than or equal to the length of *U*, besides we should get rid of the length of substitutions and deletions, so the same length and *same_rate* is defined as

$$same=length(U)-(substitutions+deletions)$$

$$same\_rate=same/len \qquad (2)$$

For instance, *S*1=*www.baduu.co*, *S*2=*www.baidu.com*, the process to calculate distances is like Figure 2.

The different distance between *S*1 and *S*2 is 3, as it changes for 3 times, once for deletion, the other two for insertion. *diff= Leven(S*1*,S*2*)=3, diff_rate*=3/11. And the same distance is 9, as the same number of characters is 9, *same_rate*=9/11.

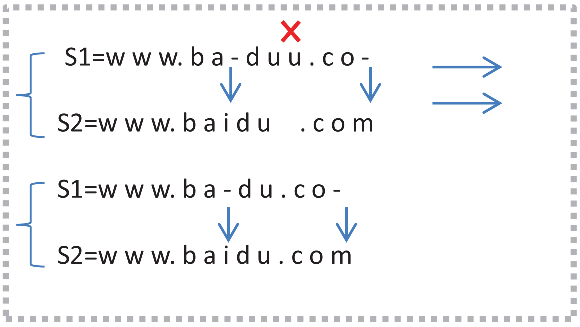
S1=w w w. b a - d u u . c o -
S2=w w w. b a i d u   . c o m
S1=w w w. b a - d u . c o -
S2=w w w. b a i d u . c o m

Figure 2 Process to calculate the distances in two domains

3.3 Define the best matching URL

The greater the similarity with sites in the Vulnerable Sites List, the smaller in different distance, the more likely it is a malicious web site. In order to find the imitation sites, we should iterate over all of the sites in the Vulnerable Sites List. Suppose the number of URLs in the Vulnerable Sites List list is w, the *i* th URL is $L_{i.}$ and the URLs in the Vulnerable Sites List is L, where $L=\{L_1,L_2, \dots , L_w\}$. the best math URLs in the Vulnerable Sites List should satisfy the different distances are as small as possible, the same distances are as large as possible. The objective function is like below.

$$\begin{cases} \min(\frac{Leven(U,L_i)}{length(L_i)}) \\ \max(\frac{same}{length(L_i)}) \end{cases}, i = 1,2, \dots , w \qquad (3)$$

As the multi-targets programming is hard to solve in theory, we introduce the method of weighting, and set the weighting values of the two objects are *a* and *b*. Finally, the function is simplified as follow.

$$f(x) = min(a*diff\_rate-b*same\_rate) \qquad (4)$$

We choose 20000 URLs to calculate the accuracy through choosing different combination *a* and *b*, and the results are as follow.

Table I Accuracy in different arguments

(*a*,*b*)	Accuracy
(0.2,0.8)	0.895
(0.4,0.6)	0.875
(0.5,0.5)	0.885
(0.6,0.4)	0.905
(0.8,0.2)	0.9

Finally, we can see that, when we choose *a*=0.6,*b*=0.4 as the combination , we will get the best answer.

3.4 Distribution in rate between phishing domains and benign domains

Any a domain input will be compared with the Vulnerable Sites List and find the best one with the minimum difference and maximum similarity, and put the distance with them as the final result.

We find some phishing sites and benign sites, and compute URL Correlation, the results are like that.

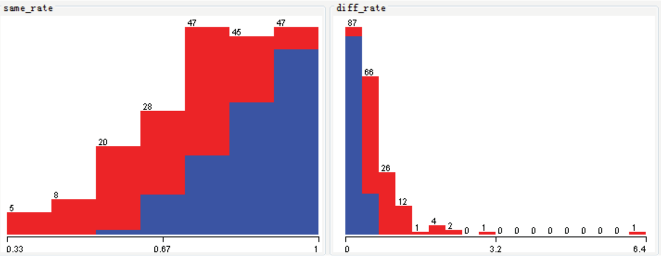

Figure 3 diff_rate and same_rate

In the picture above, the horizontal axis shows the numerical value of rate and the vertical axis shows the number of abscissa. While, red Histograms describe the phishing URLs and the blue one are the benign URLs. From the Figure 3 we can see that the same_rate in the phishing site is less and the diff_rate is bigger than the blue one, which are more related with the URLs in the Vulnerable Sites List.

4 Experimental Analysis

We collect benign URLs from DMOZ. The fishing URLs were collected from the blacklist of malicious sites. This datasets include 10000 benign URLs and 10000 phishing URLs. We collect 240 fishing URLs and 240 benign URLs to get the creation and the other information of WHOIS like [7].

To decide the stand or fall of the algorithm, five indexes are as follow.

(1)True Positive, TP: correctly identified values;

(2)True Negative, TN: correctly rejected values;

(3)False Negative, FN: incorrectly rejected values;

(4)False Positive, FP: incorrectly identified values;

(5)AVG, AVG=(TP+TN)/2: correctly average values.

In order to improve the accuracy, we used the10 -fold cross-validation [9].

We calculate the Lexical and Host Based features in 20000 kinds of URLs, which include the Lexical and host based features, the length, symbols , the number of capital letter and so on which we mentioned in Section 2. As shown in figure 4, the horizontal axis shows the numerical value and the vertical axis shows the number of abscissa. Blue represents the benign URLs, red represents the phishing sites. Part feature number comparison chart is like below.

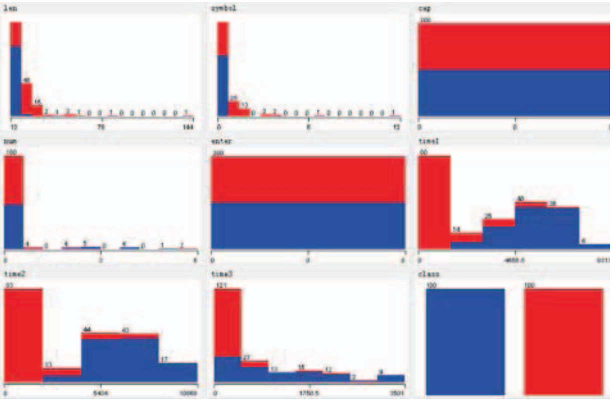

Figure 4 Distribution of features in the number of length, special symbol, capital letter, numeric scale, carriage return, time1,time2 and time3.etc.

The results reveal that there are obvious differences in the numbers of URLs lengths, the symbols, the Time in phishing sites and benign URLs.

At last, we compare the accuracy of SVM(Support Vector Machine), RBF Network and method in reference [5] added to our new feature (URL correlation) with the others just use the normal feature. The processes we do are like that

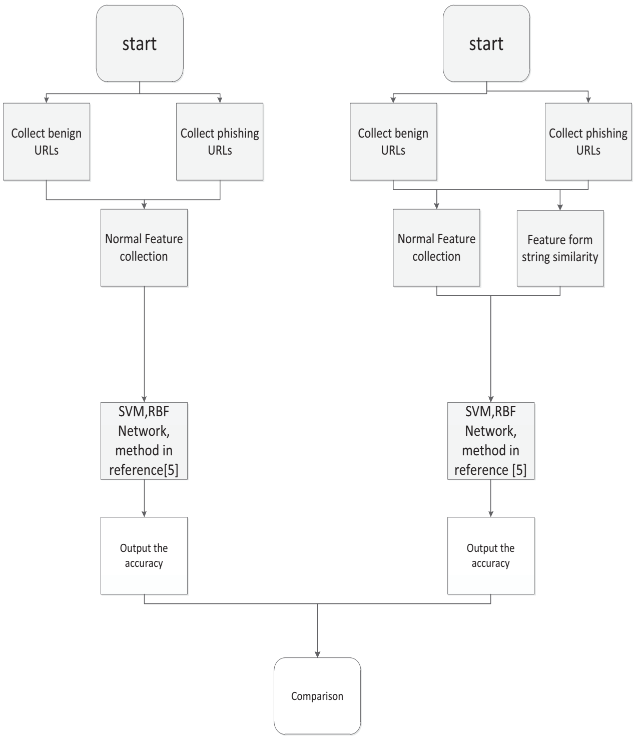
start
Collect benign URLs
Collect phishing URLs
Normal Feature collection
SVM,RBF Network, method in reference[5]
Output the accuracy
start
Collect benign URLs
Collect phishing URLs
Normal Feature collection
Feature form string similarity
SVM,RBF Network, method in reference [5]
Output the accuracy
Comparison

Figure 5 Process we do in this paper

We take URL Correlation as the new feature and add it to the normal feature. And we compare the accuracy with SVM(Support Vector Machine) and RBF Network and method in reference [5] between new feature before and after use. From the table and feature below, we can find that the accuracy has been improved, when the new feature is introduced. The comparison is like below.

Table II Comparison in 3 kinds of machine learning methods

Accu r-acy	(a)	(a+)	(b)	(b+)	(c)	(c+)
TP	0.96	0.97	0.97	0.97	0.97	0.98
TN	0.95	0.96	0.94	0.97	0.94	0.96
FN	0.04	0.03	0.03	0.03	0.03	0.02
FP	0.05	0.04	0.06	0.01	0.06	0.04
AVG	0.955	0.965	0.955	0.98	0.955	0.97

(a)Support Vector Machine(SVM) (a+) Improved SVM (b)RBF Network(b+) Improved RBF Network [(c)method used in reference [5] (c+) Improved method used in reference[5]

(a), (b), (c) utilize the commonly used feature, (a+), (b+), (c+) added the feature which we put forward. Finally, we can find that the accuracy has been improved after the new features are added, especially when we add our new feature to the method of RBP Network. And our method is better than method in reference [5].

5 Conclusions

In order to classify phishing URLs in a higher accuracy, many researchers improved their personal ideas. In the aspect of feature collections, Ma et al [5] proposed the lexical and host-based features. Joby James et al [7] came up with page importance, Chen Zhuang et al [6] raised a new method WOE to put the information of registration authority into collection features.

In this paper, a new feature URL Correlation, which includes the different rates and same rates is proposed. In order to find the best matching URL, we introduce the method of weighting and put the correlation with the matching URL as the new feature. At last, we compared the accuracy in new features with the others without through machine learning methods SVM, RBF Network and also compared with method in reference [5]. Finally, we can find our feature improved the accuracy of the classification.

Acknowledgments

The author thanks the editor and reviewers for their suggestions to improve the quality of paper.

This work was supported by the NSF of China (U1433105), NSF of China (U1536118), and NSF of China (U1536116), Beijing Higher Education Young Elite Teacher Project (YETP0448).

References

[1] China's information security published by Rising Antivrus in 2014. [OL]. http://www.rising.com.cn/about/news/rising/2015-01-07/16930.html. 2015.

[2] Hou Y T, Chang Y, Chen T,et al. Malicious web content

detection by machine learning. Expert Systemswith Applications, 2010(01): 55-60.

[3] Wang Haifeng, Duan Youxiang. Virus detection based on behavior analysis engine improvement research. Computer applications. 2004(24):109-110.

[4] LiYang, LIUBiao, FENG Hua-min. Malicious WebPages Detection Based on Machine Learning. Journal of Beijing Electronin Science and Technology Institute. 2012(4):36-40.

[5] J. Ma, L.K. Saul, S. Savage, G.M. Voelker. Beyond Blacklists: Learning to detect malicious web sites from suspicious URLs, In: Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009:1245-1254.

[6] Chen Zhuang, Liu Longfei. Malicious sites detection based on the information of registration authority. Computer CD software and application. 2015(1).

[7] JobyJames, Sandhya,L, Ciza,Thomas. Detection of Phishing URLs Using Machine Learning Techniques. International Conference on Control Communication and Computing (ICCC). 2013.

[8] http://www.alexa.com/topsites. 2016-04-19.

[9] Whittaker, C., Ryner, B., and Nazif, M., Large-Scale Automatic Classification of Phishing Pages. in NDSS'10 Proceedings of the NDSS Symposium 2010, San Diego, California, USA, 2010. DOI= http://www.internetsociety.org/doc/large-scale-automatic classification-phishing-pages.

[10] Blum, A, Wardman, B, Solorio, T., and Warner, G, http://www.internetsociety.org/doc/large-scale-automatic classification-phishing-page.

[11] Mohammed Nazim Feroz,Susan Mengel. Phishing URL detection using URL Ranking. IEEE International Congress on Big Data.2015.

[12] S. Garera, N. Provos, M. Chew, A.D. Rubin. A framework for detection and measurement of phishing attacks. Proc. 5th ACM Workshop on Recurring Malcode.

[13] Cao Jiuxin, Dong Dan, Mao Bo, Wang Tianfeng. Phishing detection method based on URL features. Journal of Southeast University (English Edition) Nanjing, China. 2013, Vol.29,No2,pp.134-138.