Detecting Phishing Web sites: A Heuristic URL-Based Approach

Luong Anh Tuan Nguyen†, Ba Lam To†, Huu Khuong Nguyen† and Minh Hoang Nguyen*
† Faculty of Information Technology, Ho Chi Minh City University of Transport
* Faculty of Management Information System, University of Economics and Law, VNU–HCM
Email: {nlatuan, balam, nhkhuong}@hcmutrans.edu.vn, nmhoang@gmail.com

***Abstract*—With the growth of Internet, e-commerce plays a vital role in the society. As a result, phishing, the act of stealing personal user data used in e-commerce transaction, has been becoming an emergency problem in modern society. Many techniques have been proposed to protect online users, e.g. blacklist, pagerank. However, the numbers of victims have been increasing due to inefficient protection technique. This is due to the fact that phishers try to make the URL of phishing sites look similar to original sites. In this paper, we are interested in proposing a new approach to detect phishing site by using the features of URL. Particularly, we derive different components from URL and compute a metric for each component. Then, the page ranking will be combined with the achieved metrics to decide whether the websites are phishing websites. The proposed phishing detection technique was evaluated with the dataset of contains 9,661 phishing websites and 1,000 legitimate websites. The results show that our proposed technique can detect over 97% phishing websites.**

I. Introduction

In recent years, e-commerce transactions have been growing rapidly due to the growth of online shopping. Buying products is easier than ever with the aid of e-commerce industry. As a consequence, e-commerce has been received more attention by the hackers who try to exploit vulnerabilities for illegitimate benefits. The most popular vulnerability of e-commerce is phishing. The term phishing is achieved by replacing the letter 'f' in the word fishing with 'ph' to make it a new word. Specifically, Phishing is that the attackers try to persuading the online users with faked e-mails or websites in order to steal their personal information, e.g. personal identification number, password, credit card numbers, etc. According to [1], phishing is defined as: "Phishing is an e-mail fraud method in which the perpetrator sends out legitimate-looking email in an attempt to gather personal and financial information from recipients. Typically, the messages appear to come from well known and trustworthy websites".

With these activities of phishing, it causes severe economy loss all over the world. For example, [2] shows that phishing attacks caused $3.2 billion loss in 2007 in United States. Therefore, it is envisaged that the phishing detection techniques must be applied in e-commerce in order to enhance the performance and the quality of the transactions and services. Many techniques have been proposed to detect the phishing websites. Generally, mainstream anti-phishing methods utilize either URL blacklists or features using machine learning algorithms. While the former is simply based on the URL matching which has weak resistance to the growth of phishing websites, the latter is really complicated and URL plays a minor role in detecting the phishing websites.

As we know that phishers can not use the exact URL which they are targeting, they must use different deceptions to build domain names and paths that look similar to the original domain. It means that URL might have great impact on whether the websites are phished. Thus, in this paper, we focus on URL to detect phishing websites. Specifically, many features of URL will be studied and considered as heuristics so that we can use them to detect whether the websites are phished. Moreover, the pagerank is also included in the heuristic set to achieve higher accuracy detection level.

The remainder of the paper is organized as follows. The related works are discussed in Section II. Section III presents the design and implementations of the heuristic URL-Based phishing detection technique. In section IV, the numerical results of experiment are illustrated. Finally, Section V concludes this paper and figures out the future works.

II. Related Works

Although many techniques have been proposed to detect phishing, they can be classified into three categories[3]: blacklist, heuristics and machine learning. In the first category, the phishing detection techniques[4][5][6][7][8] maintain a list of phishing websites called blacklist. When a URL is submitted, it will be compared to each entry in the blacklist. If there exists a match, the URL will be concluded to be the phishing URL. However, this approach is inefficient because of the rapid growth in the number of phishing websites. With heuristic approach[9][10], a signature database of known attacks will be built and used by antiviral systems or intrusion detection systems to scan a web page. The websites will be considered as phishing websites if the heuristic patterns of the websites match signatures in the database. Unfortunately, the main drawback of this approach is that signatures are easily bypassed by attackers (mainly through obfuscation) and hence, the heuristics fail to detect novel attacks. Beside that, the updating rate of the signature database is usually slower than the pace which attackers overwhelm victims with novel attacks, resulting in zero-day exploits. Similarly, the machine learning approach[11][12][13] exploits many characteristics of the URL and the websites by using machine learning techniques. These characteristics are combined to use to detect the phishing

websites. However, there are some limitations in this approach. First, the machine learning-based techniques might fail in the case that attackers compromise legitimate domains and host phishing attacks on those servers. Second, because of text-based analysis mechanism, these phishing detection techniques can not detect the phishing websites which are purely made up of images. Unlike the previous proposals in which the URL has a minor role in detecting phishing, our approach focuses on the URL to detect phishing websites. Besides, ranking of a websites (PageRank, AlexaRank, AlexaReputation) were also studied and included in our proposed technique to enhance the accuracy of the Phishing website detection.

III. System Design

A. URL

A URL (uniform resource locator) is used to locate the resources. An example of a typical URL would be "$http : //en.example.org/wiki/Main_Page$". Technically, a URL is a type of uniform resource identifier (URI). However, in many cases, URL is often used as a synonym for URI [14].

The structure of URL is as follows:

$< protocol >: // < subdomain > . < primarydomain > . < TLD > / < pathdomain >$

Each component of the URL is considered as a feature. For example, with the URL: *http://www.ebay.login.abc.net/login/web/index.html*, we will have six features as follows: Protocol is *http*, Subdomain is *ebay.login*, Primarydomain is *abc*, TLD is *net*, Domain is *abc.net*, Pathdomain is *login/web/index.html*

B. Features of URL

Since the phishing technique aims to fool the online users by making a fake URL which is similar to the URL of the original website, the domain-related features of the URL can be used to detect the phished websites. Specifically, PrimaryDomain, SubDomain and PathDomain of the URL are investigated to conclude the websites.

- PrimaryDomain: The phishers often create the phishing URL which looks similar to the legitimate sites in order to fool the users. For example, the URL www.paypall.com looks similar to well known website www.paypal.com. The phishers can not use the original primarydomain since it is already registered by the original company. Instead, the phishers register misspellings or similar looking of the original primarydomain.
- SubDomain: Because the owner of domain can add an arbitrary number of subdomains, the phishers often prepend the domain of the phished websites to his website. For example, prepending the subdomain "paypal.com" to any other domain may fool users into thinking that they are on the domain "paypal.com".
- PathDomain: If the phishers do not use primarydomain or subdomain to fool users, the phishers can use the pathdomain to fool users. The pathdomain is a subfolder of the URL. For example, with the URL http://www.attack.com/paypal, if users are not careful, they will think that they are on the "paypal" site.

C. Features of Domain's Ranking

Besides, it is obvious that the phished websites are neither accessed by the users or linked by the other websites. Therefore, the frequency that the websites are accessed by users or linked by other websites can be used as heuristic parameters to decide whether the websites are phishing. As the result, PageRank, AlexaRank and AlexaReputation are used to capture these access and linkage frequencies.

- PageRank: PageRank is a link analysis algorithm and is used by the Google web search engine. Most phishing websites' PageRank value are low because these sites exist only for a short time. Hence, PageRank value is used to classify whether a page is phishing page. Note that PageRank value for phishing site will be low and for legitimate site its value will be high.
- AlexaRank: Alexa is now the most prestigious site about the statistical information of website traffic. Alexa provides the rank of website based on traffic information, access levels, links to other websites and the updated information. Most phishing websites exist only for a short time so its AlexaRank value is very low due to our experiment with dataset from PhishTank. We use AlexaRank value to classify whether a page is phishing page. Similar to PageRank, AlexaRank value will be low in case of phishing websites and high in case of legitimate websites. In other words, AlexaRank value also plays a key role in detecting phishing sites.
- AlexaReputation: AlexaReputation value is calculated as the number of links which are linked to your website from other websites. We also use AlexaReputation value to classify whether a page is phishing page. Similar to AlexaRank and PageRank, AlexaReputation value will be high in case of phishing websites and low in case of legitimate websites.

D. System Model Design

The system model can be depicted in Figure 1.

1) Phase I - Receiving URL: URL is received from dataset or browser.

2) Phase II - Selecting four features of URL: In this phase, the features are selected from URL such as Domain, PrimaryDomain, SubDomain and PathDomain. After that, they are separated into different components. For example, if the URL is "http://www.paypal.login.abc.net/login/web/index.html", the features of URL will be: *abc* (PrimaryDomain); *paypal.login* (SubDomain); *login, web* and *index.html* (PathDomain); *abc.net* (Domain).

3) Phase III - Calculating six values of the heuristics: In this phase, we built a whitelist that contains the PrimaryDomain of the legitimate website as shown in Table I. If the value of each heuristic is negative, the site is suspected as the phishing site. If each value is positive, the site is considered as the legitimate site.

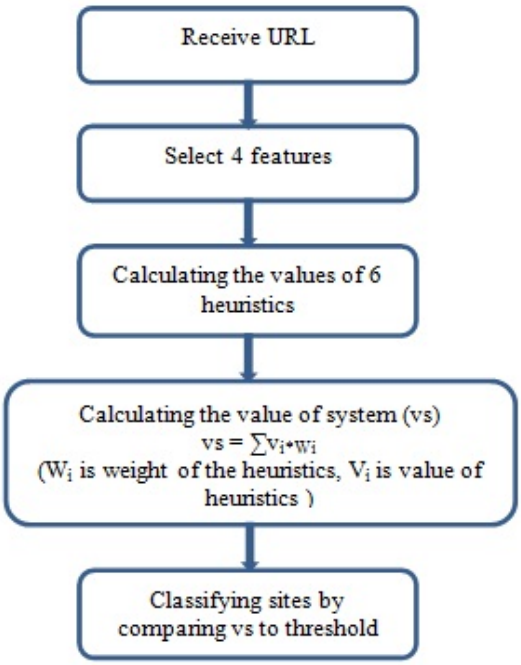

Fig. 1. The System Model

TABLE I
WHITE LIST

ID	Target
1	Ebay
2	Paybal
3	Yahoo
4	Gmail
…	...

```
Input: PrimaryDomain
Output: Value of heuristic "PrimaryDomain"

If PrimaryDomain is null then value=-0.5; //Domain is IP
If  PrimaryDomain is not null then
result = Suggestion(PrimaryDomain);
If  result is null then
        value = -0.25;
End if
If result is not null then
        Test = Check result in whitelist;
        If Test is yes then
            length=Levenshtein(result, PrimaryDomain);
           if (length = 0 )
                value = 1;
           end if
           if (length>0 and length<3)
                value = -0.5;
           end if
           if (length >= 3 )
                value = 0.25;
           end if
        Else
           length= Levenshtein(result, SubDomain);
           if (length=0)
                value = 0.5;
           end if
           if (length>0 and length<3)
                value = -0.25;
           end if
           if (length>=3)
                value = 0.25;
           end if
        End if
End if
End if
```

Fig. 2. Calculating the value of the heuristic "PrimaryDomain"

- Calculating the value of heuristic "PrimaryDomain": Based on experimental results of 9,661 phishing sites achieved by using the classifier, the algorithm used to calculate the value of the heuristic "PrimaryDomain" is shown in Figure 2.
- Calculating the value of heuristic "SubDomain": Based on experimental results of 9,661 phishing sites achieved by using the classifier, the algorithm used to calculate the value of the heuristic "SubDomain" is shown in Figure 3.
- Calculating the value of heuristic "PathDomain": Based on experimental results of 9,661 phishing sites achieved by using the classifier, the algorithm used to calculate the value of the heuristic "PathDomain" is shown in Figure 4.
- Calculating the value of heuristic "PageRank": The Google's PageRank value can be obtained from [15]. Its value varies from -1 to 10. Based on experimental results of 9,661 phishing sites achieved by using the classifier, we propose the algorithm to calculate the value of heuristic "PageRank" (Figure 5).
- Calculating the value of heuristic "AlexaRank": AlexaRank value can be obtained from [16]. Based on experimental results of 9,661 phishing sites achieved by using the classifier, the algorithm used to calculate the value of the heuristic "AlexaRank" is shown in Figure 6.
- Calculating the value of heuristic "AlexaReputation": AlexaReputation value can be obtained from [16]. Based on experimental results of 9,661 phishing sites achieved by using the classifier, the algorithm used to calculate the value of the heuristic "AlexaReputation" is shown in Figure 7.

4) Phase IV - Calculating the value of system (vs): The value of the system is calculated by (1)

$$vs = \sum v_i * w_i \quad (1)$$

Where v_i is value of each heuristic, w_i is weight of each heuristic. The weight of each heuristics can be obtained by using the classifier through the result of experiment based on dataset 9,661 phishing sites from PhishTank. Specifically, the weights are given in the Table II.

```
Input: SubDomain
Output: Value of heuristic "SubDomain"

If SubDomain is null then value=0;
If  SubDomain is not null then
     result = Suggestion(SubDomain);
     If  result is null then
          value = 1;
     End if
     If result is not null then
          Test = Check result in whitelist;
          If Test is yes then
                value = -1;
          Else
                length= Levenshtein(result, SubDomain);
                if (length==0)
                      value = -0.5;
                end if
                if (length>0 and length<3)
                      value = -0.25;
                end if
                if (length>=3)
                      value = 0.5;
                end if
          End if
     End if
End if
```

Fig. 3. Calculating the value of the heuristic "SubDomain"

```
Input: PathDomain
Output: Value of heuristic "PathDomain"

If PathDomain is null then value=0;
If PathDomain is not null then
     result = Suggestion(PathDomain);
     If  result is null then
          value = 1;
     End if
     If result is not null then
          Test = Check result in whitelist;
          If Test is yes then
                value = -1;
          Else
                length= Levenshtein(result, PathDomain);
                if (length==0)
                      value = -0.5;
                end if
                if (length>0 and length<3)
                      value = -0.25;
                end if
                if (length>=3)
                      value = 0.5;
                end if
          End if
     End if
End if
```

Fig. 4. Calculating the value of the heuristic "PathDomain"

5) Classifying the websites: In this phase, we compare the value of "vs" to threshold to decide whether a website is a phishing website. Figure 8 shows the algorithm.

```
Input: PageRank value
Output: Value of heuristic "PageRank"

If PageRank value <=0 then value = -1
If PageRank value in [1..2]  then value = -0.5
If PageRank value in [3..4] then value = -0.25
If PageRank value in [5..6]  then value = 0.25
If PageRank value in [7..8] value = 0.5
If PageRank value in [9..10] then value = 1
```

Fig. 5. Calculating the value of the heuristic "PageRank"

```
Input: AlexaRank value
Output: Value of heuristic "AlexaRank"

If AlexaRank value >= 3.000.000 then value = -1
If AlexaRank value in [2.000.000 .. 3.000.000] then value = -0.5
If AlexaRank value in [1.000.000 .. 2.000.000] then value = -0.25
If AlexaRank value in [500.000 .. 1.000.000] then value = 0.25
If AlexaRank value in [300.000 .. 500.000] then value = 0.5
If AlexaRank value < 300.000 then value = 1
```

Fig. 6. Calculating the value of the heuristic "AlexaRank"

```
Input: AlexaReputation value
Output: Value of heuristic "AlexaReputation"

If AlexaReputation value <5 then value = -1
If AlexaReputation value in [5..10] then value = -0.5
If AlexaReputation value in [11..15] then value = -0.25
If AlexaReputation value in [16..20]  then value = 0.25
If AlexaReputation value [21..30] value = 0.5
If AlexaReputation value >30 then value = 1
```

Fig. 7. Calculating the value of the heuristic "AlexaReputation"

```
If  vs < thresold then alert "phishing site"
Else alert "original site"
End if
```

Fig. 8. Classifying the websites

IV. Evaluation

We have selected 9,661 phishing sites from PhishTank[4] as training dataset and testing dataset that contains 1,000 legitimate sites from DMOZ[17] and 1,000 phishing sites from PhishTank. Experimental procedure is divided into 5 phases.

A. Phase 1

Dataset is collected from PhishTank[4] and imported into MYSQL with 9,661 phishing websites. The result is shown in the Figure 9.

B. Phase 2

Four features (PrimaryDomain, SubDomain, PathDomain and Domain) are selected. In this phase, we use PHP to select four features as PrimaryDomain, SubDomain, PathDomain and Domain from URLs in dataset. Figure 10 shows the obtained result.

phish_id	url	phish_detail_url	submission_time	target
1777670	http://www.bishopflahertyassembly.org/paypal/paypa...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-04-01 16:16:41	PayPal
1777669	http://www.usenmeztente.com.tr/media/patrick/DarkM...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-04-01 16:15:28	PayPal
1777668	http://paypal.ausadia.com/1fbeefecbb20b2cee2160061...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-04-01 16:15:17	PayPal
1777666	http://paypal.ausadia.com/ae439d217f67cbb7bd1c4da0...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-04-01 16:15:16	PayPal
1777662	http://www.dawhoo.com/wp-content/docfile/fileindex...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-04-01 16:14:10	AOL

Fig. 9. Dataset of 9,661 phishing sites in MYSQL

phish_id	domain	primarydomain	subdomain	pathdomain
1777670	bishopflahertyassembly.org	bishopflahertyassembly		paypal,paypal,priv8,us,2012,pp,login.htm
1777669	usenmeztente.com.tr	usenmeztente		media,patrick,DarkMalier
1777668	ausadia.com	ausadia	paypal	
1777666	ausadia.com	ausadia	paypal	
1777662	dawhoo.com	dawhoo		wp,content,docfile

Fig. 10. Four features are extracted

TABLE II
WEIGHT OF THE HEURISTICS

Heuristic	Phishing sites	Weight
PrimaryDomain	2971	0.105
SubDomain	1380	0.049
PathDomain	3787	0.134
PageRank	7514	0.266
AlexaRank	6437	0.227
AlexaReputation	6208	0.219

TABLE III
RESULT OF CLASSIFYING WEBSITES

Threshold	Phishing sites	Ratio
−0.2	4830	50%
−0.1	5120	53%
0	6122	63%
0.1	6768	70%
0.2	7539	78%
0.3	8696	90%
0.4	9182	95%
0.5	9374	97%
0.6	9661	100%

C. Phase 3

In this phase, search engine spelling suggestions and alexa.com are used to calculate the value of the heuristics. The result is shown in the Figure 11.

D. Phase 4

In this phase, we calculate the value of "vs" for each URL from dataset of 9,661 phishing sites and compare to the thresholds. The results are shown in Table III.

E. Phase 5

In this phase, our proposed technique is tested with testing dataset which contains 1,000 phishing sites from PhishTank and 1,000 legitimate sites from DMOZ[17]. In case of 2-class prediction, there are four possible results that are defined as follows:

- True Positive (TP): If the result of prediction is legitimate site and the actual value is also legitimate site.
- False Positive (FP): If the result of prediction is phishing site but the actual value is legitimate site.
- True Negative (TN): If the result of prediction is phishing site and the actual value is also phishing site.
- False Negative (FN): If the result of prediction is legitimate site but the actual value is phishing site.

If the accuracy ratio is calculated as follows: Accuracy_ ratio = (TP+TN)/(TP+TN+FP+FN), the results of the test will be shown in Table IV.

TABLE IV
RESULT OF TESTING

Threshold	TP	TN	FP	FN	Accuracy Ratio
−0.2	623	497	377	503	56%
−0.1	576	602	424	398	59%
0	611	627	389	373	62%
0.1	587	634	413	366	61%
0.2	675	705	325	295	69%
0.3	795	868	205	132	83%
0.4	889	945	111	55	92%
0.5	979	967	21	33	97%
0.6	876	997	124	3	94%

From the obtained results, we have found that this technique has a high accuracy rate of 97% with the threshold value of 0.5.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a new technique to detect phishing sites effectively in real time. In the pro-

phish_id	pagerank	alexarank	alexarep	primarydomain	subdomain	pathdomain	vs
1777670	-1	1	-1	-0.25	0	-1	-0.17675
1777669	-1	1	-1	0.25	0	0	-0.03525
1777668	-1	1	-1	0.25	-1	0	-0.08425
1777666	-1	1	-1	0.25	-1	0	-0.08425
1777662	-0.5	1	-0.25	0.25	0	0	0.09775

Fig. 11. Values of heuristics

posed technique, we used six heuristics (primarydomain, subdomain, pathdomain, pagerank, alexarank, alexareputation) whose value and weight are calculated from the experimental results. The training data set that contains 9,661 phishing sites. We have collected the testing data set that contains 1,000 legitimate and 1,000 phishing sites to test with 97% accuracy. In the future, we will classify the value and the weight of each heuristic more accurate with many training datasets. Furthermore, more heuristic parameters will be added, such as WHOIS properties, Domain name properties, Geographic properties [18] and combined with machine learning technique to achieve a hybrid technique.

REFERENCES

[1] SecuritySearch. (2006, October). [Online]. Available: http://searchsecuritytechtarget.com/definition/phishing

[2] McCall. Gartner survey. [Online]. Available: http://www.gartner.com/it/page.jsp?id=565125

[3] M. Khonji, A. Jones, and Y. Iraqi, "A novel phishing classification based on url features," in *IEEE GCC Conference and Exhibition*, Dubai, 2011, pp. 221–224.

[4] PhishTank. (2013, April) Statistics about phishing activity and phishtank usage. [Online]. Available: http://www.phishtank.com/stats/2013/01/

[5] G. D, *Google bots detect 9,500 new malicious websites every day*. http://arstechnica.com/security/2012/06/google-detects-9500-newmalicious-websites-daily/, 2012.

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang. (2009) An empirical analysis of phishing blacklists. [Online]. Available: http://ceas.cc/2009/papers/ceas2009-paper-32.pdf

[7] Google. (2011, August) Google safe browsing api. [Online]. Available: http://code.google.com/apis/safebrowsing/

[8] McAfee. (2011, July) Mcafee site advisor. [Online]. Available: http://www.siteadvisor.com

[9] C. Seifert, I. Welch, and P. Komisarczuk, "Identication of malicious web pages with static heuristics," in *The Australasian Telecommunication Networks and Applications Conference*, 2008.

[10] N. Chou, R. Ledesma, Y. Teraguchi, and J. Mitchell, "Client-side defense against web-based identity theft," in *The 11th Annual Network and Distributed System Security Symposium*, 2004. [Online]. Available: http://crypto.stanford.edu/SpoofGuard/webspoof.pdf

[11] P. Likarish, D. Dunbar, and T. E. Hansen, "B-apt: Bayesian anti-phishing toolbar," in *IEEE International Conference on Communications*, May 2008, pp. 1745–1749.

[12] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in *The 16th international conference on World Wide Web*, 2007, pp. 639–648.

[13] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: a feature-rich machine learning framework for detecting phishing web sites," *ACM Transactions on Information and System Security*, vol. 14, no. 2, pp. 1–28, Sept. 2011.

[14] Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Uniform_resource_locator

[15] G. Inc. [Online]. Available: http://toolbarqueries.google.com

[16] Alexa. [Online]. Available: http://data.alexa.com/data?cli=10&dat=snbamz&url=

[17] DMOZ. (2013). [Online]. Available: http://rdf.dmoz.org/rdf/

[18] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklist: Learning to detect malicious web sites from suspicious urls," in *The 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France., June 28 - July 1 2009, pp. 1245–1254.