THE URL FEATURES FOR PHISHING BY USING WORD SUGGESTION

1I UON-CHANG LIN, 2HUNG-CHIEH CHUANG

1Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan
2Department of Photonics and Communication Engineering, Asia University, Taichung, Taiwan
E-mail: iclin@nchu.edu.tw

Abstract- In recent years, with the rapid development of technology, internet has become a tool we often use, it brings us convenience for the daily life, and personal information has become increasingly important. It also results in number of cyber-attacks, among them, phishing attacks are often used to steal information. There are different method to forge a website, due to the different time and environment. In order to detect phishing attacks effectively, the goal of this study propose some new features to prevent phishing attacks. We first analyze the URL before users loading sites, check the word combinations whether is meaningful or not. We obtain features and classify the website according to these features. It is determined whether the site is a phishing site. Overall, we propose the method in the paper can identify their phishing sites effectively, the method of analysis URL with 95.71 % accuracy rate.

Keywords- Phishing, Classification, Features

I. INTRODUCTION

With the development of the Internet, there are variety of services in the Internet. It can handle a variety of things in life. The Internet brings convenience, at the same time, the value of personal information also rise. It followed many threats, therefore it result personal data be stolen, among the internet threats, phishing attack is often used to steal personal information in recent years. Phishing websites means phishers imitation of a well-known websites, and deceive the user to obtain the user's confidential information. If the users does not pay more attention for the website, they cannot distinguish the difference between legitimate sites and phishing sites.

Although the relevant research has proposed many detection methods, but still not able to reduce the threat of phishing effectively, phishing sites are still rapidly generated. Phishers in order to lure user fall into the trap, will try a new way to avoid detection current phishing tools [1]. The phisher often send e-mail to the victims, if victims do not careful of the message, it will think that the mail is sent from a legitimate website and click the links in the mail. Victims will be directed to fakes sites, entering personal information to phishing website, the phisher will steal the victim's data, and victim will fall into trap accidentally. If the users can analyze the URL before they click it, then they might not avoid phishing attack.

Therefore, in this paper, we propose the features for detecting phishing websites for analysis the URL. We observe the difference between legitimate sites and phishing website.
II. RELATED WORK

According to the timing of user browsing the websites, it can be divided into two types, URL analysis, and web content analysis. The URL analysis is check the URL before user load the website and analyzed it. The content analysis is user enter the website and download the data to analyze in local.

PhishNet [2] believes that phishers often through simple URL changes, to avoid comparison, first use heuristics of known phishing sites to find new phishing websites, then apply approximate matching algorithm to generate one final score. If the score is greater than threshold, the website will be consider a potential phishing site. In [3], it combine whitelist and SVM classifier, when the URL has a high similarity with the whitelist URLs, but it does not the white list, it will be judged as phishing. If there is the same URL in the list, then recognized as a legitimate site. Low similarity will use an SVM to make predictions based on some features of phishing. There is a research [4] focuses on the similarity of phishing site’s URL and legitimate site’s URL. It use Levenshtein Distance [5] to calculate the similarity between legitimate and phishing site URLs, and combined website ranking, calculated weights to classify phishing sites.

In CANTINA [6], it is a content-based approach to detect phishing, besides the URL and its domain name basically, CANTINA use TF-IDF algorithm to retrieve information. TF-IDF can measure how important a word in a document, and from their experiments result, it is good at detecting phishing, and using TF-IDF can reduce FP rate effectively. MobiFish [1] use OCR to detect content in mobile page, turn into text, and then compare the current URL. If the URL of the current page does not include the conversion of text through OCR, it will be considered as a phishing site. In 2014, a study [7] through detect the webpage between direct and indirect links, and applying Target Identification(TID) algorithm to detect phishing website and find the target which is mimicked.

III. BACKGROUND

URL (Uniform resource locator) means that you need key some word in browser that you want to enter a website. It is shown in Fig 1. When you key the URL in a browser, it will get the content of the website from the Internet and show it on your computer finally.

![Fig 1. URL](image)

The URL is the only way to connect website on the Internet, no matter you click a link in e-mail or in a webpage. If we can pay more attention to the URL, we can effectively avoid phishing attack or other Internet threats.

For example:

http://mail.google.com

It can be divided into five parts:

1. Scheme
2. Domain Name
3. Sub domain
4. Primary domain
5. TLD

In the above example, http is scheme, mail is sub domain, google is primary domain, and com is TLD.

When a company needs to build a new website, it has to register a domain name of its own domain, usually use its own brand to represent its own unique website. No one can use the same domain except the domain name registrant. Because the phishers cannot use same
domain, they only can use similar primary domain name to confuse the users. In addition to the primary domain, the company may provide a number of web service, they use a sub domain to direct users to specific web services, such as mail.google.com, it means that a mail service is provided by Google. The text of primary domain can represent their own unique sites. It usually will be recognized by everyone, the sub domains also can be understand what user browse current site. In other words, the composition of the URL will be meaningful.

In order to check whether the URL is meaningful or not, it uses the Google search engine and Fig 2 shows the result from Google.

![Google Suggest](image)

**Fig 2. Google Suggest**

When we try to find something on the search engine, the search engine will list suggestion associated with input. If the keyword is the current popular search words then the search engine will be able to show the user wants accurately. The search suggestion provided by the system based on various algorithms [8], it can help us to find out the meaningful text. In addition, other search engines have similar functions, it shown in Fig 3.

![Yahoo Search Suggestion](image)

**Fig 3. Yahoo Search Suggestion**

Then explain the way generate features, analyzing the composition of URL is the fastest and most direct way. The URL on behalf of an image of company, usually use its brand name be a domain name, the sub domain is other service. Therefore, when we obtain the URL, it will keep domain only and separated it by dot.

Because TLD has global uniform format on the web, we removed gTLD (Generic Top-Level Domains) first, and ccTLD (Country-Code Top-Level Domains) after leaving the word are primary domain and sub domain from the back to front.

For example, https://mail.google.com is the URL of Gmail. First keep the words between "//" and "/", get mail.google.com, and then separated it by dot, get mail, google, com. Next remove TLD, then get mail and google, so google is primary domain and mail is sub domain.

**IV. EXPERIMENTS**

![Methodology](image)

**Fig 4. Methodology**

The process is shown in Fig 4. When it receives URL, the URL will split into three parts based dot. They are
extracted from URL such as domain, primary domain, and sub domain. Second, which will key the words in the search engine, and then get the suggestion from each word. For example, we search facebook on search engine. It well return <facebook, facebook login, facebook search…>. Third, if the words we input are in the suggestion list, it means the word is a meaningful word. If the word is not a meaningful word, the suggestion will return null. Finally it generates the value according the return results.

Feature 1: Domain
1. The URL is too long, phishing sites will join the text of legitimate website in the URL for confuse users. It may use many words to mimic.
2. The words of domain appeared in the list of search suggestion. It means that is a popular domain.
3. The words of domain do not appeared in the list of search suggestion. It means that is a phishing domain.

Feature 2: Sub Domain
1. It is too long.
2. There is not sub domain.
3. It is appeared in the list of search suggestion
4. It is not appeared in the list of search suggestion

Feature 3: Primary Domain
1. It is appeared in the list of search suggestion
2. It is not appeared in the list of search suggestion

In the experiments, we select 1000 phishing sites from PhishTank [9], 165 legitimate sites are from Alexa [10] including the legitimate sites with other services domain.

We mainly used JAVA to perform the analysis, the data all stored in MYSQL. It divides into 80% of the training data, 20% of test data for classification for classification. The experiments use LIBSVM [11] to classify phishing, it provides a simple way to use and get accurate results.

<table>
<thead>
<tr>
<th>Classified as phishing</th>
<th>Classified as legitimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>phishing</td>
<td>199</td>
</tr>
<tr>
<td>legitimate</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>24</td>
</tr>
</tbody>
</table>

From Table 1, it can calculate the Accuracy of 95.7082%. The method can identify phishing effectively through keyword search. As the results, there are nine website cannot classify to legitimate in legitimate websites. The fact from test dataset, if the sub domain uses abbreviation such as mobile abbreviated as m, then the system will return search suggestions of m. If m is not in list, it may return maps, not m only. Thus, it results in incorrect classification. In the results of phishing sites, only one website is classified as legitimate sites, it means the method can identifying phishing in most cases. In fact, the website will classify as phishing correctly if it analyze from URL only. The website may be hacked and it is placed in a phishing page, that resulting incorrect classification.

CONCLUSION
In this paper, it proposed URL-based approach to analyzing phishing sites. The scope of application including e-mail, website, and text messages on mobile. It is provide a quick method to analyze before user enter the website. It requires text of URL only, and it does not cost too many resources.

For users, whether user browse the site or check E-mail, if you need click link, the method can provide a quick way to analyze. It can help users to filter most
phishing, protecting users from receive spam or phishing, further to protect the users safety and increasing intention to use internet.

We proposed a quick and effective way to identify phishing sites. It is a method through search suggestion, which use three features. In the experiments, we collects 1,000 phishing websites and 165 legitimate site, which the method can effectively classify phishing sites, the accuracy is 95%. In the future, the search engines can provide different result according different region, which can use the ccTLD information to get local search suggestion.

REFERENCE


***