DETECTION OF PHISHING PAGES OVER THE INTERNET USING ANTI-PHISHING TECHNIQUES

Prof P Jayalakshmi[1], Prof M N Rajaprabha[2], Prof R Vijay Anand[3]

[1] *School of Information Technology & Engineering, VIT (India)*

[2] *School of Information Technology & Engineering, VIT (India)*

[3] *School of Information Technology & Engineering, VIT (India)*

ABSTRACT

Phishing can be defined as an attempt used by hackers or intruders to obtain sensitive information about the users like passwords, usernames, and id’s and so on to gain access and steal information without the user’s knowledge. This usually occurs by email-spoofing malware infected sites that imitate the original websites. In this era of growing technology, a great number of the population is influenced by the internet and social media. And phishing has always been a great threat to the people. To overcome this shortcoming, we propose an idea to employ an anti-phishing technique, “similarity detection of HTML source code” using an anti-phishing browser to compare the textual content between the authentic and phishing pages.

Index terms: Phishing, spoofing, malware, hacking, anti-phishing.

I INTRODUCTION

Phishing can be understood as a technique or an attempt to of sending a false email to a user in order to acquire his user identity and gain access over his privacy. Sometimes these emails are sentto thousands of people by the phisher through which he keeps track on the number of emails being read by the ratio of people, who apparently have an account with a legitimate enterprise or company. This type of spoofing is also called as brand spoofing and carding. For instance in the year 2003, the internet users faced a lot of increase in the phishing scam. One such was a spoof email from ebay.com where it claimed that the user’s account will be deactivated if they do not update their credit card credentials in a specified link given in the email. The Phisher could easily imitate the original website by just using the HTML code. And people were fooled of being suspended. Similar scenarios led to the need for developing Anti-Phishing software[1].

The Anti-phishing software encompasses of programs that identify phishing content in emails and websites which are usually incorporated with browsers and clients to prevent others masquerading as a legitimate website. This functionality can also be integrated as a built-in capability of web browsers. Several anti-phishing techniques have been used in the recent times like Content-Based Approach, Envelope Heuristics, Reputation Systems, Visual similarity based Approach, Blacklist based, Webpage Characteristics based solutions, and Webpage Context based Solutions [2].

II LITERATURE SURVEY

Roopak.S et.al has proposed a method for detecting phishing pages [3]. They have used a technique of web mining to search similar web pages through Google and used their HTML source codes to compare them. If nothing matched with the HTML source code then Cosine similarities of those results are computed to identify the phishing pages by developing their own browser.

More than 20 phishing sites are being tested with that browser along with HTML source code and cosine similarity values. It is being implemented with java with jsoup html parser and gson package. Once a match is found, it is known that the pages are similar and those IP address are checked. Difference in IP address represents the phishing page else a legitimate page. Their experimental analysis shows better results in detecting phishing sites than existing methods.

M.Madhuri et.al has proposed a work on phishing detection and prevention systems [4]. In this paper the authors have introduced a technique called linkGuard which is the new end-host based anti-phishing algorithm to make use of the standard characteristics of the hyperlinks that are common in phishing attacks. The difference between the actual link and visual links are analyzed by linkGuard and also determines the similarities of a URL with a known authenticated site.

It is been implemented by windows XP and common standard characteristics of those hyperlinks have helped linkGuard to detect known as well as unknown phishing attacks. Their experiments shows effective linkGuard with minimum false negatives in detecting known phishing attacks along with unknown phishing attacks. They have also showed that linkGuard is seems to be light-weighted process and it has detected 96%of unknown phishing attacks.

Engin Kirda et.al have proposed a work on Protecting Users Against Phishing Attacks [5]. They have used a browser extension called AntiPhish to protect the users from phishing attacks. They have taken Mozilla browser extension and are written in JavaScript and XML User Interface Language. It tracks the user information like password, social security number, etc and gives warning whenever the user gives the sensitive information to a web site which is considered to be untrusted.For storing sensitive information about the user, JavaScript DES implementation is used.

After the user enters his confidential information such as password the menu called AntiPhish is selected and it scans the page to capture and store the information. Also the current contents of HTML fields can also be retrieved. Apart of information it can also store the source. In this they have used domain instead of web site addresses because in some are hosted by multiple servers. They also provide simple dialogs for stored information and also the users can view the list of domains and also can clear the information.

III COMPARISON OF EXISTING TECNIQUES

S.No.	Phishing Technique	Description	Merits	Demerits
1.	Content-Based Approach [2]	Content-based filtering is proved to be the effective method in discovering the unnecessary content and also identifies many phishing attacks. It functions on the message content to evaluate keywords and phrases of concern. It assigns an overall probability for a specificgroup.	• Does not provide false positive and • Does not produce zero day phishing	• Categories must be predefined. • Not suitable for identifying specific targeted organizations. • Filtering efficiency is lower. • No proof of copying only provides hints.
2.	Visual similarity based Approach[2]	An approach to detect phishing web pages that supports visual similarity, which may be used by a legitimate webpage owner look for suspicious webpages that are visually similar.	• Dynamic image generation • Finds similar web pages over the internet • Calculates similarity measures • Threshold comparison.	• Time-consuming • Low accuracy rate • More time to calculate pair of pages • Difficult to measure similarity.
3.	The Detection System based on URLs Features[2]	Most common anti-phishing technique used today. Uses many different learning algorithms.	• Easy to compare between legitimate and phishing websites. • Analyses target website characteristics	• No general internet features • Differences in URL feature due to wide range of domains in various regions.

4.	Blacklist based[6, 7, 8]	Phishing is one of the easiest and effective ways for trickery over the internet. However solutions like URL blacklisting are effective to certain extent in detecting phishing	• Straightforward for attackers to elude. • Simple implementation by the browsers and many other different applications.	• Unity • Tools might have false negatives • No timely detection of threats
5.	Phishing detection based on behaviour of user[9]	It is a new technique to detect phishing websites by analysing the online users' behaviours like visited websites and the data used. These behaviours cannot be manipulated by phishers and achieves high accuracy during detection.	• Detects essential characteristics of a phishing attack. • The information of how the users can be manipulated may change with future phishing attacks. • Greater detection accuracy. • Analyses the identity of websites using IP addresses and domain names.	• It cannot handle all types of authentication details. • It handles static credentials like username; password etc, but dynamic credentials cannot be handled (e.g. OTP).

IV ANTI-PHISHING TECHNIQUE

Our system employs an anti phishing technique that searches similar web pages through Google and compare the HTML source code of these pages. Suppose, there is no similarities or match in the code then the pages are compared using the cosine similarities. The working of our method is shown in Figure 1.

A) Drawing URL through Google Search

In this step, a signature is formed by making use of the web page title and few words of it with the highest Term Frequency (TF) values. These signatures are then supplied in Google search and first n web pages are retrieved. Then, the URL supplied by the user is compared with these results to check for similarities.

B) Comparison of web pages

In this step HTML code is matched between the web pages and also the textual content is checked through cosine similarity. If either of these process results in matches then those pages are alleged to be similar. So we consider two types of web pages, primary and secondary web pages. Where the URL supplied webpage is called primary and returned web page results through search are called secondary web pages. So, these primary pages are compared with first n secondary pages, if there prevails a match then, IP addresses of those pages are checked. If the addresses are same then they are genuine pages otherwise phishing pages.

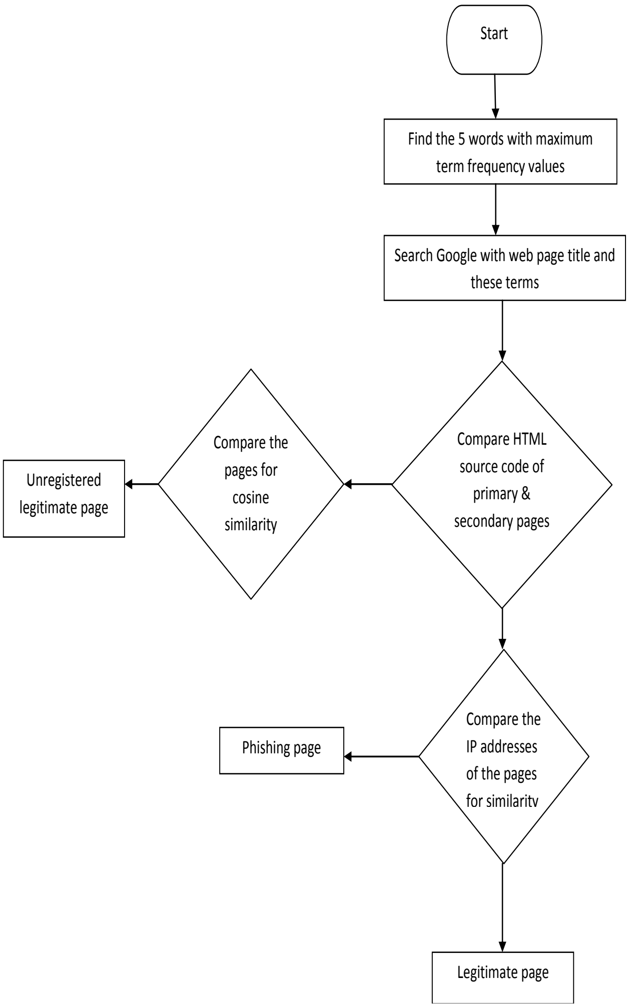

Start
Find the 5 words with maximum term frequency values
Search Google with web page title and these terms
Compare HTML source code of primary & secondary pages
Compare the pages for cosine similarity
Unregistered legitimate page
Compare the IP addresses of the pages for similarity
Phishing page
Legitimate page

Fig.1 Workflow

C) Comparison of HTML source code

In this step, the HTML tags of the primary and secondary pages are compared using attribute matching. The number of tag mismatches and matches of the pages and the percentage of the matches are calculated. If the value exceeds 50% then the pages are similar. The algorithm for this process is given below.

```
**Data:** Vector M, Vector N
**Result:** match_found
a=0;
b=0;
match=0;
```

```
while M(a) is empty do
    while N(b) is empty do

        if M(a)==N(b) then
            if M(a).attribute == N(b).attribute then
                match=match+1;
                a=a+1;
            else
                b=b+1;
            end
        else
            b=b+1;
        end
    end
    a=a+1;
end
if ((match/.totaltags)*100) is greater than 50 then
    match_found=true;
else
    match_found=false;
end
```

D) Cosine Similarity Comparison

The textual content of the primary and secondary web pages are compared for cosine similarity. If its value surpasses 0.50, then the pages are considered similar.

E) Implementation

We have created a webpage that takes up a URL supplied by the user and validates whether the entered URL is a legitimate one or a Phishing page. If the entered URL is a valid one then the corresponding web page will be displayed. Otherwise, the web page will generate a warning message that it is a phishing web page.

We have taken some of the phishing URL's from Phishtank.com [10] to verify our system. The screenshots of our system are shown below. Figure 2, depicts the browser that we have created called the "Mini Browser" for validating the URL.

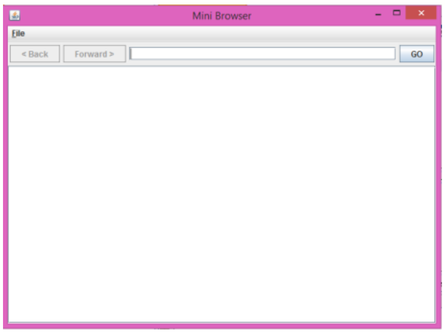
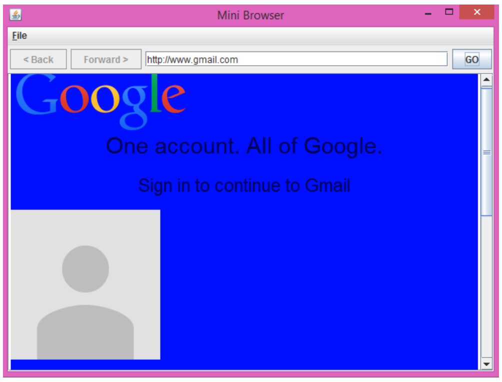


Fig. 2 Browser page

Figure 3 depicts the validation of a genuine web page.



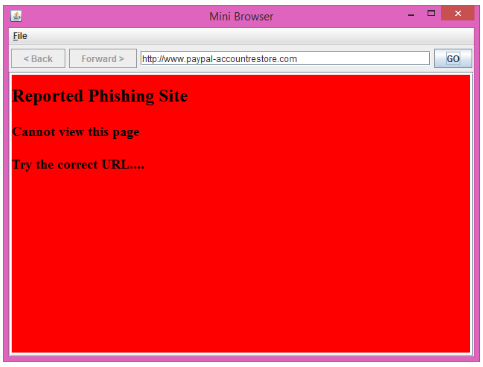
http://www.gmail.com

One account. All of Google.

Sign in to continue to Gmail

Fig. 3 Genuine page

Figure 4 depicts the validation of a phishing page and display of a warning message.



http://www.paypal-accountrestore.com

Reported Phishing Site

Cannot view this page

Try the correct URL.....

Fig. 4 Warning message

Thus, through our created browser we detect the phishing pages depending of the similarity frequencies and prompt the user with a warning message to prevent them from accessing a phishing page. Through this we can witness a good security mechanism where the users can be benefited immensely.

V CONCLUSION

In the above study, we conclude that most of the Anti-Phishing techniques focus on contents of web page, URL and email. Character Based Anti-Phishing approach may result in false positive but Content Based Approach never results in false positive. Attribute Based Approach consider almost all major areas vulnerable to phishing so it can be best anti-phishing approach that can detect known and unknown phishing attack. Our system detects the URL and performs similarity function to identify Phising attack.

As a future work on phishing we can do more work on server side security. In the server side security policy we use dual level of authentication for user by which only authentic user can get the access of his account, and to educate the user about this policy will results in avoiding user to give his sensitive information to phished web site.

In this paper we have discussed a novel process to detect phishing pages. Though there is a low false positive rate and high detection rate while comparing with existing techniques, it just deals with only html source code. Since it is difficult for the humans to understand the source code easily it may not work better in case if any code is obfuscated. This may be avoided by including the browser screen shot process in the future.

REFERENCES

[1]http://www.webopedia.com/TERM/P/phishing.html

[2] Varsha Mishra, “A survey of various anti-phishing techniques”, Universe of Emerging Technologies and Science ISSN: 2349 - 655X Volume I Issue I – June 2014.

[3] Roopak.S, Tony Thomas, “A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity”, Fourth International Conference on Advances in Computing and Communications, 2014.

[4] M.Madhuri, K.Yeseswini, U. Vidya Sagar, “Intelligent phishing website detection and prevention system by using link guard algorithm”, International Journal of Communication Network Security, ISSN: 2231 – 1882, Volume-2, Issue-2, 2013.

[5] Engin Kirda and Christopher Kruegel, “Protecting Users Against Phishing Attacks with AntiPhish”,

[6] Sujata Garera, Niels Provos, Monica Chew, Aviel D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks". Proceedings of 2007 ACM Workshop On Recurring Malcode, pp. 1-8, 2007.

[7] Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker, "Identifying Suspicious URLs: An Application of Large-Scale Online Learning". Proceedings of the 26th Annual International Conference on Machine Learning, pp. 681-688, 2009.

[8] Andre Bergholz, Gerhard Paab, Frank Reichartz, Siehyun Strobel, Jeong-Ho Chang, "Improved Phishing Detection using Model-Based Features". Proceedings of the Conference on Email and Anti-Spam, 2008.

[9] Xun Dong, John A. Clark,Jeremy L. Jacob “User Behaviour Based Phishing Websites Detection”, University of York York, United Kingdom.

10] http://www.phishtank.com