ISSN No:-2456-2165

A Study on Phish Detection Methods and a Combined Approach of List, Image and Content Based Methods.

Tiny Molly V
IT Department, Viswajyothi College of Engineering,Vazhakulam, Kerala

Abstract:- In this paper various approaches towards detection of phishing sites are considered. The various phishing detection methods available are list based, image based, content based etc. The aim is to detect phishing site in an easy way. Here a combined approach of list, image and content based method is implemented. If webpage is not in list then move towards content based or image based method and check contents. Content based approach can be applied if web page contains only text and no images. If webpage lacks contents ie webpage contains only images then apply image based method.

Keywords:- *phishing, list based, image based, content based.*

I. INTRODUCTION

- Phishing is a web based attack.

Web based attacks tries to capture personal information from users. The main use of this information is in fraudulent activities. Most phishing sites are similar to an original site in its appearance so it is not easy for a user to identify between them. Phishing attacks as well as the victims are increasing day-by-day as attackers develop new ways to fool users. List based methods are the easiest way to detect phishing sites but they are not able to detect phishing sites that came into existence within a short span of time.

II. DETECTION METHODS

List based method is a phishing detection method. List based methods are classified into two. They are black list methods and white list methods. List based methods use a database to detect phishing site.

- *List Based Methods*

Most of the Internet browsers use blacklist method to detect phishing sites. In Blacklist method details of all known phishing sites are stored in a database. But white list method maintains a list of all legitimate sites. The accuracy of list based method depends on how often database is updated. So in order to identify phishing sites that came into existence within a short span of time heuristic based method is used.[3][12]

- *Heuristic Based Methods*

Heuristic-based approaches depends up on certain characteristics to detect phishing. These characteristics can be any web page characteristics such as URL, HTML code, image or the page content itself.[7]

III. RELATED WORK

CATINA a heuristic based approach was proposed by Zhang et al. Cantina uses a lexical signature heuristic based on the TF-IDF (term frequency/inverse document frequency) algorithm to detect phishing sites.[2]

Dunlop et al. proposed a heuristics-based approach called Gold Phish to detect zero day phishing sites. This method captures image of a web page and after removing ads, using OCR conversion method image is converted to text. Then Google Page Rank is applied to check whether it is a phishing site. GoldPhish achieved a phishing detection rate of 98% [1].

Prakash et al. proposed PhishNet a predictive blacklist method to detect phishing sites. It uses URL features to detect phishing attacks. [3].

B-APT a Bayesian Antiphishing Toolbar proposed by Likarish et al. for detection of phishing sites uses DOM analyzer [4].

The above mentioned approaches include list, text and image based methods. Here the above methods are combined in an effective manner.

IV. DESIGN

The proposed approach after studying list based, image based and content based method uses a combination of these three. First use blacklist method to decide whether a site is phishing site or not. If it is in blacklist then it is a phishing site and the process can be stopped. But the phishing sites that are new not necessarily be in blacklist. If the site is not in blacklist check webpage contains images/text.

List based approach can be implemented easily. Only a checking is needed. But it is not accurate. Accuracy can be maintained by frequently updating database. If list method fails and web page contains images only then image based approach is used. Even though it is accurate it is time consuming. So when list based method fails try to apply content based method as far as possible.

ISSN No:-2456-2165

If webpage contains text only then content based approach is easy to implement than image based approach. Extract text from webpage using javascript code. There is a Javascript method document. get Element By Id to gather text and store it in a variable. The content stored in the variable is applied to a google search engine. If webpage is legitimate then its name appears in first few searches.[10][11]

To extract all contents of a webpage use get Element By Tag name method, For example if x is a string use javascript code as

- *x+=document. get Element By Tag Name('*')*

This method stores all elements into variable x and contents of x are used to decide Google Page Rank. If name of webpage is within top 10 sites that are displayed in google search then it is a legitimate site. Otherwise it is a phishing site.[13]

If webpage contains links or tables or forms only they can also be collected. For this approach DOM collections property such as links collection, images collection etc are used.[11] These collections can be stored in a variable and it can also be used for google Page Rank search.

If webpage contains images only then image based method is used. First convert image to text using OCR tool. Then text can be checked using Google Page Rank mechanism. If it is not in phishing site then name of webpage will available within first few searches.[9].

Steps:

1. Obtain URL of a webpage from address bar.
2. Check URL in phishing site.
3. If found it is a phishing site.
4. Otherwise check webpage contains images only. If so apply Gold Phish page Rank mechanism.
5. If webpage contains text then it is extracted and apply Google Page Rank mechanism. Same is the case of links, form, tables etc.

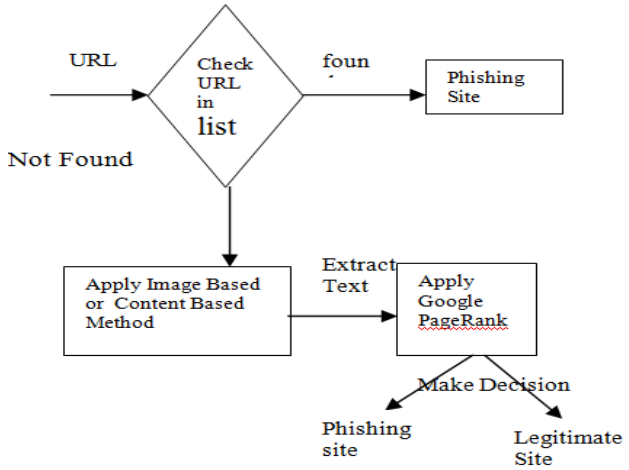

Fig 1:- Phish Detection System

V. CONCLUSION

In this paper it is tried to overcome the problems of blacklist and Gold Phish method to detect phishing sites. In blacklist method although it is faster it cannot detect newly arrived phishing sites. To overcome this problem we are integrating image based method with blacklist method. But if a webpage contains only text then image based method also fails. Here content based approach can be implemented. [13]

This combined approach is a good method but some problems may arise due to limitations in webpage. Problems are mainly due to lack of contents such as images to verify a web site. The probability of such a page is very low. In normal cases the main page contains sufficient data to verify the legitimacy of the web site. Despite these limitations this method is a faster and powerful tool for detecting zero day phishing sites.

REFERENCES

[1]. Dunlop. M., Groat. S., Shelly. D. (2010) 'GoldPhish: Using Images for Content-Based Phishing Analysis', Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference, pp: 123 – 128.
[2]. Y. Zhang, J. I. Hong, and L. F. Cranor. (2007) 'Cantina: a content-based approach to detecting phishing web sites', Proceedings of the 16th International conference on World Wide Web, New York, NY, USA, 2007. ACM, pp 639–648.
[3]. Prakash. P, Kumar. M, Kompella. R. R, Gupta. M. (2010) 'PhishNet: Predictive Blacklisting to Detect Phishing Attacks ',INFOCOM, 2010 Proceedings IEEE , pp: 1 – 5.
[4]. Likarish. P, Eunjin Jung, Dunbar. D, Hansen. T. E, Hourcade. J. P (2008) ' B-APT: Bayesian Anti-Phishing Toolbar' ICC '08. IEEE International Conference, pp: 1745 – 1749.
[5]. Y. Cao, W. Han, and Y. Le. Anti-phishing based on automated individual white-list. In DIM '08: Proceedings of the 4th ACM workshop on Digital identity management, pages 51–60, New York, NY, USA, 2008. ACM.
[6]. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
[7]. Jaydeep Solanki, Rupesh G. Vaishnav (2016) 'Website Phishing Detection using Heuristic Based Approach' International Research Journal of Engineering and Technology (IRJET),2016.
[8]. www.electronics-manufacturers.com/ .../optical-character-recognition-ocr.html.
[9]. www.google.com/support/websearch/bin/answer.py?hl=en&answer=106318.
[10]. www.w3schools.com.

[11]. Paul J. Deitel, Harvey M. Deitel, Abbey Deitel, "Internet and World Wide Web How To Program", 5/E, Pearson Education, 2012.
[12]. A.Belabed, E.Aimeur, A. chikh A Personalized Whitelist Approach for Phishing Webpage Detection, '
[13]. https://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-webmasters-11068