International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Detection of Phishing Website Using Machine Learning

Hemali Sampat[1], Manisha Saharkar[2], Ajay Pandey [3], Hezal Lopes[4]

[1,2,,3,4] *Department of Computer Engineering, Universal College of Engineering, Vasai, Maharashtra, India*

Abstract - *Detection of Phishing website is an intelligent and effective model that is based on using classification or association Data Mining algorithms. These Algorithms were used to identify and characterize all rules and factors in order to classify the phishing website and relationship that correlate them with each other so we detect them by their performance, accuracy, number of rules generated and speed. Proposed system implements both algorithms which is Classification and Association that optimizes the system which is more efficient and faster than existing system. By using these two algorithms with WHOIS protocol the error rate of the existing system decreases by 30% so by using this method proposed system create an efficient way to detect the phishing website. Although there does not exist a system which can detect all the phishing website but using these methods it will create a most efficient way to detect the phishing website.*

Key Words: **Phishing Websites, Data Mining algorithm, Association algorithm, classification algorithm, WHOIS PROTOCOL**

1. INTRODUCTION

Social engineering attack is a common security threat used to reveal private and confidential information by simply tricking the users without being detected. The main purpose of this attack is to gain sensitive information such as username, password and account numbers. According to, phishing or web spoofing technique is one example of social engineering attack. Phishing attack may appear in many types of communication forms such as messaging, SMS, VOIP and fraudster emails. Users commonly have many user accounts on various websites including social network, email and also accounts for banking. Therefore, the innocent web users are the most vulnerable targets towards this attack since the fact that most people are unaware of their valuable information, which helps to make this attack successful.

Typically phishing attack exploits the social engineering to lure the victim through sending a spoofed link by redirecting the victim to a fake web page. The spoofed link is placed on the popular web pages or sent via email to the victim. The fake webpage is created similar to the legitimate webpage. Thus, rather than directing the victim request to the real web server, it will be directed to the attacker server.

The current solutions of antivirus, firewall and designated software do not fully prevent the web spoofing attack. The implementation of Secure Socket Layer (SSL) and digital certificate (CA) also does not protect the web user against such attack. In web spoofing attack, the attacker diverts the request to fake web server. In fact, a certain type of SSL and CA can be forged while everything appears to be legitimate. According to, secure browsing connection does virtually nothing to protect the users especially from the attackers that have knowledge on how the "secure" connections actually work. This paper develops an anti-web spoofing solution based on inspecting the URLs of fake web pages. This solution developed series of steps to check characteristics of websites Uniform Resources Locators (URLs). URLs of a phishing webpage typically have some unique characteristics that make it different from the URLs of the legitimate web page. Thus, URL is used in this paper to determine the location of the resource in computer networks.

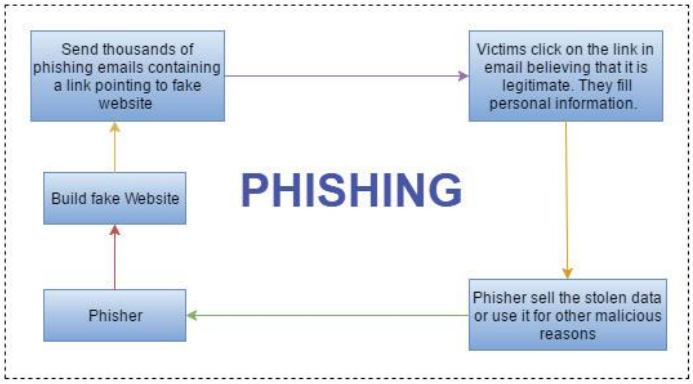


Fig. 1: Flow of general phishing attack

2. LITERATURE SURVEY

In [1] JAIN MAO, WENQIAN TIAN and ZHENKAI LIANG has proposed a system which detect the phishing using page component similarity which analyzes URL tokens to increase prediction accuracy phishing pages typically keep its CSS style similar to their target pages. Based on the observation, a straightforward approach to detect phishing pages is to compare all CSS rules of two web pages, It prototyped Phishing-Alarm as an extension to the Google Chrome browser and demonstrated its effectiveness in evaluation using real-world phishing samples.

ZOU FUTAI, PEI BEI and PAN LI [2] Uses Graph Mining technique for web Phishing Detection. It can detect some potential phishing which can't be detected by URL analysis. It utilize the visiting relation between user and website. To get dataset from the real traffic of a Large ISP. After anonymizing these data, they have cleansing dataset and each record includes eight fields: User node number (AD), User SRC IP(SRC-IP) access time (TS), Visiting URL (URL), Reference URL(REF), User Agent(UA), access server IP (DST-IP), User cookie (cookie). For a client user, he is assigned a

unique AD but a variable IP selected from ISP own IP pool. Therefore, we build the visiting relation graph with AD and URL, called AD-URL Graph and the Phishing website is detected through the Mutual behavior of the graph

In [3] NICK WILLIAMS and SHUJUN LI proposed a system which analysis ACT-R cognitive behavior architecture model. Simulate the cognitive processes involved in judging the validity of a representative webpage based primarily around the characteristics of the HTTPS padlock security indicator. ACT-R possesses strong capabilities which map well onto the phishing use case and that further work to more fully represent the range of human security knowledge and behaviors in an ACT-R model could lead to improved insights into how best to combine technical and human defenses to reduce the risk to users from phishing attacks

XIN MEI CHOO, KANG LENG CHIEW and NADIANATRA MUSA [4] this system is based on utilizing support vector machine to perform the classification. This method will extract and form the feature set for a webpage. It uses a SVM machine as a classifier which has two phase training phase and testing phase during training phase it extracts feature set and while testing it predict the website is legitimate or a phishing.

In [5] GIOVANNI ARMANO, SAMUEL MARCHAL and N.ASOKAN proposed a use of add on in the browser which is Real-Time Client-Side Phishing Prevention. It uses information extracted from website visited by the user to detect if it is a phish and warn the user. It also determines the target of the phish and offers to redirect the user there. A warning message is displayed in the foreground while the background displays the phishing webpage darkened by a black semi-transparent layer preventing interactions with the website.

TRUPATI KUMBHARE and SANTOSH CHOBE

[6] have discussed various Association Rule Mining Algorithm. Association rule learning searches for relationships among variables. Various Association algorithm discussed are AIS algorithm, SETM algorithm, Apriori algorithm, Aprioritid algorithm, Apriorihybrid algorithm, and FP-growth algorithm.

In [7] S.NEELAMEGAM and DR.E.RAMARAJ discussed various Classification Algorithm used in data mining. Data Classification is a data mining technique used to predict group membership for data instances Various Classification Algorithm discussed are decision tree, Bayesian networks, k-nearest neighbor classifier, Neural Network, Support vector machine.

VARSHARANI RAMDAS, V.Y. KULKARNI and R.A.RANE[8] proposed a system to detect a phishing website using Novel Algorithm This detection algorithm can find out the maximum number of phishing URLs because it executes multiple tests such as Blacklist search Test, Alexa ranking test, and different URL features test. But this solution is effective only for HTTP URLs.

In [9]JUN HU,YUCHUN JI and HANBING YAN

This method to detect Phishing website is based on the analysis of legitimate website server log information. every time a victim opens the phishing website, the phishing website will refer to the legal website by asking for resources. Then, there will be a log, which is recorded by the legitimate website server and from this logs Phishing site can be Detected

SAMUEL NARCHAL, GIOVANNI ARMANO and NIDHI SINGH[10] propose a application Off-the-Hook application for detection of phishing website. Off-the-Hook, exhibits several notable properties including high accuracy, brand-independence and good language-independence, speed of decision, resilience to dynamic phish and resilience to evolution in phishing techniques.

3. PROPOSED SYSTEM

This section describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites features, Blacklist and WHOIS database. According to few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity, security & encryption, source code, page style and contents, web address bar and social human factor. This study focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IP Address, long URL address, adding a prefix or suffix, redirecting using the symbol "//", and URLs having the symbol "@".These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites.

A.URL based

3.1. Using the IP Address

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal sensitive information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link

"http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF The Domain Part has an IP

Address → Phishing

Otherwise→ Legitimate

3.2. Long URL to Hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar.

For example:

http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

To ensure the accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

Rule: IF URLlength is ≤ 75→legitimate

otherwise→Phishing

We have been able to update this feature rule by using a method based on frequency and thus improving upon its accuracy.

3.3. Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example http://www.Confirme-paypal.com/.

Rule: IF Domain Name Part Includes

(-)Symbol → Phishing

Otherwise → Legitimate

3.4. Submitting Information to Email

Web form allows a user to submit his personal sensitive information that is directed to some server for processing. A phisher might redirect the user's information to his personal email. To that end, a server-side script language might be used such as "mail()" function in PHP. One more client-side function that might be used for this purpose is the "mailto:" function.

Rule: IF Using ""mail()\" or \"mailto:\" Function to Submit User Information" → Phishing

Otherwise → Legitimate

3.5.Using Pop-up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

Rule: IF Popup Window Contains Text

Fields→ Phishing

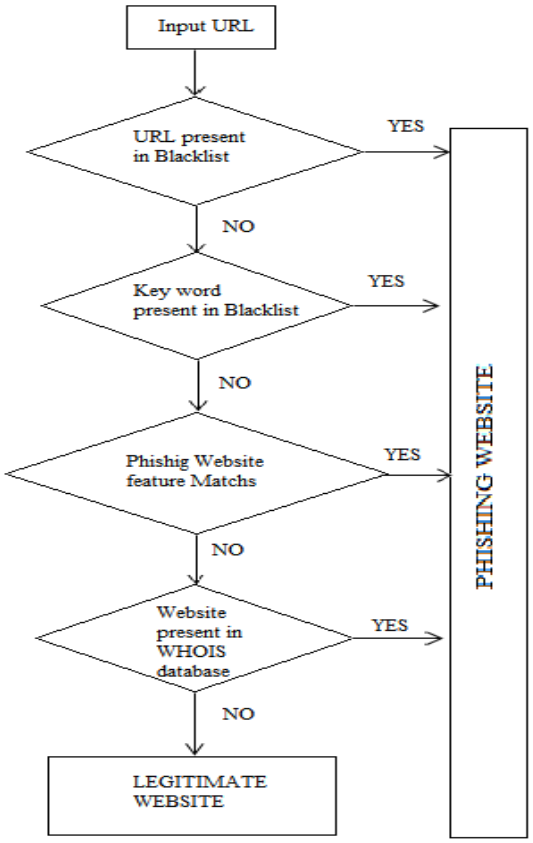
Otherwise → Legitimate

B. Blacklist based

A Blacklist is created in the proposed model in which the website detected as phishing is saved for the future use a to keep a track record and data of the phishing website this can be useful in analyzing the phishing website to increase the efficiency of the system.

C. WHOIS Database

The life of phishing site is very short, therefore; this DNS information may not be available after some time. If the DNS record is not available anywhere then the website is phishing. If the domain name of the suspicious webpage is not match with the WHOIS database record, then webpage considers as phishing.



3. CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. In Future System can upgrade to automatic Detect the web page and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. PhishChecker application also can be upgraded into the web phone application in detecting phishing on the mobile platform.

REFERENCES

[1] JIAN MAO1,WENQIAN TIAN1, PEI LI1, TAO WEI2, AND ZHENKAI LIANG3 Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity.

[2] Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen

Web Phishing Detection Based on Graph Mining.

[3] Nick Williams, Shujun Li Simulating human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behaviour architecture model.

[4] XIN MEI CHOO, KANG LENG CHIEW,DAYANG HANANI ABANG IBRAHIM,NADIANATRA MUSA, SAN NAH SZE, WEI KING TIONG FEATURE-BASED PHISHING DETECTION TECHNIQUE.

[5] Giovanni Armano, Samuel Marchal and N. Asokan Real-Time Client-Side Phishing Prevention Add-on.

[6] Trupti A. Kumbhare and Prof. Santosh V. Chobe An Overview of Association Rule Mining Algorithms.

[7] S.Neelamegam, Dr.E.Ramaraj Classification algorithm in Data mining: An Overview

[8] Varsharani Ramdas Hawanna, V. Y. Kulkarni and R. A. Rane A Novel Algorithm to Detect Phishing URLs.

[9] Jun Hu, Xiangzhu Zhang,Yuchun Ji, Hanbing Yan, Li Ding, Jia Li and Huiming Meng Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs.

[10] Samuel Marchal, Giovanni Armano and Nidhi Singh Off-the-Hook: An Efficient and Usable.

[11] U.Naresh1 U.Vidya Sagar2 C.V. Madhusudan Reddy3 IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3 (Sep. - Oct. 2013), PP 28-36 www.iosrjournals.org

[12] W. D. Yu, S. Nargundkar, N. Tiruthani, "Phishcatch – a phishing detection tool", 33rd Annual IEEE International on Computer Software and Applications Conference 2009. COMPSAC '09, pp. 451-456, 2009.

[13] How to recognize phishing email messages or links", March 2011, [online] Available: http://www.microsoft.com/security/online-privacy/phishing-symptoms.aspx.

[14] P. Likarish, D. Dunbar, T. E. Hansen, "B-apt: Bayesian anti-phishing toolbar", IEEE International Conference on Communications 2008. ICC '08, pp. 1745-1749, 2008.

[15] A. Y. Fu, L. Wenyin, X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)", IEEE Trans. Dependable Secur. Comput., vol. 3, no. 4, pp. 301-311, Oct. 2006.

[16] G. Liu, B. Qiu, L. Wenyin, "Automatic detection of phishing target from phishing webpage", Pattern Recognition (ICPR) 2010 20th International Conference on, pp. 4153-4156, aug. 2010

[17] Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey" arXiv:1701.07179v2 [cs.LG] 16 Mar 2017

[18] https://mafiadoc.com/phishing-websites-features_59b4677d1723ddd9c6441fec.html

[19] http://resources.infosecinstitute.com/category/enterprise/phishing/phishing-countermeasures/anti-phishing-the-importance-of-phishing-awareness-training/

[20] P. Prakash, M. Kumar, R. R. Kompella, M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks", INFO COM'10: Proceedings of the 29th conference on Information communications. Piscataway NJ USA:IEEE Press, pp. 346-350, 2010.

[21] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, C. Zhang, "An empirical analysis of phishing blacklists", Proceedings of the 6th Conference in Email and Anti-Spam ser. CEAS'09 Mountain view CA, July 2009.

[22] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, "A comparison of machine learning techniques for phishing detection", Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit ser. eCrime'07. New York NY USA:ACM, pp. 60-69, 2007

BIOGRAPHIES

Ms. Hemali Sampat is currently pursuing **B.E** degree in Computer Engineering from the Universal College of Engineering, Vasai, Mumbai University, Mumbai, India.

Ms. Manisha Saharkar is currently pursuing **B.E** degree in Computer Engineering from the Universal College of Engineering, Vasai, Mumbai University, Mumbai, India.

Mr. Ajay Pandey is currently pursuing **B.E** degree in Computer Engineering from the Universal College of Engineering, Vasai, Mumbai University, Mumbai, India.

Mrs. Hezal Lopes is currently HOD of Computer Engineering in Universal College of Engineering, Vasai, Mumbai. She received her **M.E** in Computer Engineering from Mumbai University.