International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

Phishing Website Detection using Machine Learning : A Review

Purvi Pujara*[1], M. B.Chaudhari[2]

*[1] Student, Computer Department, Government Engineering College, Gandhinagar, Gujarat, India

[2]Professor, Computer Department, Government Engineering College, Gandhinagar, Gujarat, India

ABSTRACT

Phishing is the fraudulent attempt to obtain sensitive information such as username, password, bank account details, and credit card details for malicious use. Phishing frauds might be the most popular cybercrime used today. There are various domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was targeted by phishing more than in any other industry sector. Several anti-phishing techniques are there such as blacklist, heuristic, visual similarity and machine learning. From this, blacklist approach is commonly used because it is easy to use and implement but it fails to detect new phishing attacks. Machine Learning is efficient technique to detect phishing. It also removes drawback of existing approach. We perform detailed literature survey and proposed new approach to detect phishing website by features extraction and machine learning algorithm.

Keywords : Phishing Detection, Feature Extraction, Phishing Website, Phishing Attacks

I. INTRODUCTION

Now days, As there are so many people are being aware of using internet to perform various activities like online shopping, online bill payment ,online mobile recharge, banking transaction .Due to wide use of this customer face various security threats like cybercrime .There are many cybercrime that are widely performed for example spam , fraud ,cyber terrorisms and phishing. Among this phishing is new cybercrime and very popular nowadays. Phishing is fraud attempt, which performed to obtain sensitive information of user. Phisher design website which looks same as any legitimate site and spoof user for obtaining private information of user such as username, password, banking details for miscellaneous reasons.

According to APWG 2Q report [2].the total number of phish detected in 2Q 2018 was 233,040, compared to 263,538 in 1Q 2018. These totals exceed the 180,577 observed in 4Q 2017 and the 190,942 seen in 3Q 2017.

There were increases in SAAS/webmail targeted sector with 21% of overall phishing attack. Payment sector is continuing as most attractive target for phishing.

According to APWG 1Q report [3], the total number of phish detected in 1Q 2018 was 263,538.This was up 46 percent from the 180,577 observed in 4Q 2017. It was also significantly more than the 190,942 seen in 3Q 2017.The number of unique phishing reports submitted to APWG during 1Q 2018 was 262,704, compared to 233,613 in 4Q 2017 and 296,208 in 3Q 2017.

II. BACKGROUND THEORY

Phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords and credit card details, often for malicious reasons, by disguising as a trustworthy entity in an electronic communication [1].Phishing attack can be implemented in various form like Email phishing, Website phishing, spear phishing, Whaling, Tab napping, Evil twin phishing etc.To avoid this phishing attack various anti-phishing solutions should be use. There are various anti phishing solutions such as Blacklist, heuristic, visual similarity, machine learning etc.

A. Blacklist method

This is most commonly used approach in which list of phishing URL is stored in database and then if URL is found in database, it is known as phishing URL and gives warning otherwise it is called legitimate. This approach is easy and faster to implement as it see URL is in db or not. But limitations is small change in URL is sufficient to bypass the list based technique and Frequent update of list is necessary to counter new attack.

B. Heuristic based method

This is extension of blacklist and able to detect new attack as use features extracted from phishing site to detect phishing attack. But limitation is cannot detect all new attack and easies to bypass once attacker know algorithm or features used. In addition, this has poor detection because site may or may not have common features.

C. Visual similarity

This approach deceive user by extracting image of legitimate site. But limitation of this is image comparison takes more time as well as more space to store image .produces high false negative rate and fail to detect when visual appearance slightly changes.

D. Machine learning

This approach works efficiently in large dataset. This also removes drawback of existing approach and able to detect zero day attack .Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99% .Performance depends on size of training data, feature set, and type of classifier. Limitation of this is it fails to detect when attacker use compromised domain for hosting their site.

Many of research have been performed in this area of phishing detection. Most research has worked on improving accuracy of phishing website detection using different classifiers. Various Classifiers used are KNN, SVM, Decision tree, ANN, Naïve Bayes, PART, ELM and Random forest. Among all of this tree based classifiers DT and RF is best as increase dataset as per my literature survey. Therefore, proposed approach will be on phishing website detection using tree-based classifiers.

Various performance measure used for analysis of best algorithm are F-measure, precision, recall, accuracy, AUC, ROC curve etc.

III. LITERATURE SURVEY

Rao et al. [6] proposed a novel classification approach that use heuristic based feature extraction approach. In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party-based features, Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services. Also this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has a

limitation of more execution time for the websites with more number of links.

Chunlin et al. [7] proposed approach that primarily focus on character frequency features. In this they have combined statistical analysis of URL with machine learning technique to get result that is more accurate for classification of malicious URLs. Also they have compared six machine-learning algorithms to verify the effectiveness of proposed algorithm which gives 99.7% precision with false positive rate less than 0.4%.

Sudhanshu et al. [8] used association data mining approach. They have proposed rule based classification technique for phishing website detection. They have concluded that association classification algorithm is better than any other algorithms because of their simple rule transformation. They achieved 92.67% accuracy by extracting 16 features but this is not up to mark so proposed algorithm can be enhanced for efficient detection rate.

M. Amaad et al.[9] presented a hybrid model for classification of phishing website. In this paper, proposed model carried out in two phase. In phase 1,they individually perform classification techniques, and select the best three models based on high accuracy and other performance criteria. While in phase 2, they further combined each individual model with best three model and makes hybrid model that gives better accuracy than individual model. They achieved 97.75% accuracy on testing dataset. There is limitation of this model that it requires more time to build hybrid model.

Hossein et al.[10] developed an open-source framework known as "Fresh-Phish". For phishing websites, machine-learning data can be created using this framework. In this, they have used reduced features set and using python for building query .They build a large labelled dataset and analyse several machine-learning classifiers against this dataset .Analysis of this gives very good accuracy using machine-learning classifiers. These analyses how long time it takes to train the model.

Gupta et al. [11] proposed a novel anti phishing approach that extracts features from client-side only. Proposed approach is fast and reliable as it is not dependent on third party but it extracts features only from URL and source code. In this paper, they have achieved 99.09% of overall detection accuracy for phishing website. This paper have concluded that this approach has limitation as it can detect webpage written in HTML .Non-HTML webpage cannot detect by this approach.

Bhagyashree et al.[12] proposed a feature based approach to classify URLs as phishing and non-phishing. Various features this approach uses are lexical features, WHOIS features, Page Rank and Alexa rank and Phish Tank-based features for disguising phishing and non-phishing website. In this paper, web-mining classification is used.

Mustafa et al.[13] developed safer framework for detecting phishing website. They have extracted URL features of website and using subset based selection technique to obtain better accuracy .In this paper, author evaluated CFS subset based and content based subset selection methods And Machine learning algorithms are used for classification purpose.

Priyanka et al.[14] proposed novel approach by combining two or more algorithms. In this paper ,author has implemented two algorithm Adaline and Backpropion along with SVM for getting good detection rate and classification purpose.

Pradeepthi et al.[15] In this paper ,Author studied different classification algorithm and concluded that tree-based classifier are best and gives better accuracy for phishing URL detection. Also Author uses various

features such as lexical features, URL based feature, network based features and domain based feature.

Luong et al. [16] proposed new technique to detect phishing website. In proposed method, Author used six heuristics that are primary domain, sub domain, path domain, page rank, and alexa rank, alexa reputation whose weight and values are evaluated. This approach gives 97 % accuracy but still improvement can be done by enhancing more heuristics.

Ahmad et al.[17] proposed three new features to improve accuracy rate for phishing website detection. In this paper, Author used both type of features as commonly known and new features for classification of phishing and non-phishing site. At the end author has concluded this work can be enhanced by using this novel features with decision tree machine learning classifiers.

Mohammad et al. [18] proposed model that automatically extracts important features for phishing website detection without requiring any human intervention. Author has concluded in this paper that the process of extracting feature by their tool is much faster and reliable than any manual extraction.

Table 1 : Evaluation Metrics

$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}}$
$False\ Positive\ Rate(FPN) = \frac{\text{FP}}{\text{FP} + \text{TN}}$
$False\ Negative\ Rate(FNR) = \frac{\text{FN}}{\text{FN} + \text{TP}}$
$Accuracy = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$
$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$
$Error\ rate = \frac{\text{FP} + \text{FN}}{\text{P} + \text{N}}$

IV. CONCLUSION

Phishing is a way to obtain user's private information via email or website. As usage of internet is very vast, almost all things are available online now it is either about shopping cloths, electronic gadgets, crockery or to payment of mobile, TV & electricity bill. Rather than standing out in line for hours, people are being aware of using online method. Due to this phisher has wide scope to implement phishing scam. As there is lot of research work done in this area, there is not any single technique, which is enough to detect all types of phishing attack. As technology increases, phishing attackers using new methods day by day. This enables us to find effective classifier to detection of phishing.

In this paper, we performed detailed literature survey about phishing website detection .According to this, we can say tree-based classifiers in machine learning approach is best suitable than other.

V. REFERENCES

[1] Phishing definition, https://en.wikipedia.org/wiki/Phishing

[2] APWG Report 1, http://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf

[3] APWG report 2, http://docs.apwg.org/reports/apwg_trends_report_q1_2018.pdf

[4] Phishing dataset, https://www.phishtank.com/developer_info.php

[5] J. Han and M. Kamber, Data Mining Concepts and Techniques, Elsevier, 2006.

[6] Routhu Srinivasa Rao1 , Alwyn Roshan Pais : Detection of phishing websites using an efficient feature-based machine learning framework :In Springer 2018.

[7] Chunlin Liu, Bo Lang : Finding effective classifier for malicious URL detection : In ACM,2018

[8] Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi : Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison : In Springer,2018.

[9] M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani : A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms :In International Conference on Computational Science and Computational Intelligence IEEE ,2016.

[10] Hossein Shirazi, Kyle Haefner, Indrakshi Ray: Fresh-Phish: A Framework for Auto-Detection of Phishing Websites: In (International Conference on Information Reuse and Integration (IRI)) IEEE,2017.

[11] Ankit Kumar Jain, B. B. Gupta : Towards detection of phishing websites on client-side using machine learning based approach :In Springer Science+Business Media, LLC, part of Springer Nature 2017

[12] Bhagyashree E. Sananse, Tanuja K. Sarode : Phishing URL Detection: A Machine Learning and Web Mining-based Approach : In International Journal of Computer Applications,2015

[13] Mustafa AYDIN, Nazife BAYKAL : Feature Extraction and Classification Phishing Websites Based on URL : IEEE,2015

[14] Priyanka Singh, Yogendra P.S. Maravi, Sanjeev Sharma : Phishing Websites Detection through Supervised Learning Networks : In IEEE,2015

[15] Pradeepthi. K V and Kannan. A: Performance Study of Classification Techniques for Phishing URL Detection: In 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE,2014

[16] Luong Anh Tuan Nguyen†, Ba Lam To† ,Huu Khuong Nguyen† and Minh Hoang Nguyen : Detecting Phishing Web sites: A Heuristic URL-Based Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13)

[17] Ahmad Abunadi, Anazida Zainal ,Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach :In IEEE,2013.

[18] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey: An Assessment of Features Related to Phishing Websites using an Automated Technique:In The 7th International Conference for Internet Technology and Secured Transactions,IEEE,2012