International Conference on Recent Innovation in Engineering, Science and Management (RIESM-19)
Osmania University Centre for International Program, Osmania University Campus, Hyderabad (India)
23rd November 2019 www.conferenceworld.in
ISBN : 978-81-943584-3-5

Survey on Detection of Phishing Website Using Machine Learning

Shwetha Bhat[1], Sanjana Honnappa Nayak[2], Surabhi[3], Pooja J[4]

Students of Department of Information Science and Engineering Canara engineering college, Benjanapadavu

ABSTRACT: *This project is focused on Detection of Phishing technique using Machine Learning. Due to the rapid growth of the internet, websites have become the intruder's main target. Phishing cost internet users lots of dollars per year. It refers to exploiting weakness on the user side. An intruder embeds malicious contents in a web page for the purpose of doing some bad and unwanted - activities such as : credential information and resource theft, luring a user to visit a dangerous website, downloading and installing software to join a botnet or to participate in distributed denial of service, and even damage the visitor system. As the number of web pages increases, the malicious web pages are also increasing and the attack is increasingly become sophisticated. In this we provide a framework for detecting a malicious web page using artificial neural network learning techniques. In addition to the significant detection rate, our objective is to find also which discriminative features characterize the attack and reduce the false positive rate. The algorithm is based on two features group, the URL lexical and the page content features. The experiments has shown the expected Results and the high false positive rate which produced by machine learning approaches is reduced.*

***Keywords*— Feature Extraction; Phishing website; Phishing Detection;**

I.INTRODUCTION:

The internet has become an essential in our daily life; it's the base of banking transactions, shopping, entertainment, resource sharing, news, and social networking. The growth of the web rewarded the cyber criminals towards it, with this growth, also the design and the use of the malware scenario has changed, its more steal their and polymorphic than damaging the machines. The majority of malware is intended to either steal the user's private data, or force the victim system to join a malware distribution network. Web is a common method for spreading malware; the attackers exploit the vulnerabilities of web browsers, web application, and operating system to gain control of a victim's machine, which is used to host various malicious activities, such as heap spray, dot net, key loggers, sending spam emails, and so on. Further techniques such as PHP language, adobe flash, and visual basic script are in common have capability of download and execute code from the Internet .In addition most browsers have a feature of plug-ins,

ISBN : 978-81-943584-3-5

which allow third parties to extend the functionality of the browser.Although several solutions have been proposed to fight malicious software, but web site exploits has not received much attention so far. In this paper, we provide a framework for detecting a malicious web page based on two groups of features using artificial neural network (ANN). This work is continuing to, add some values to the field malware combat, mitigate some threats, and improve Performance by enhancing the detectionrate. Machine Learning is efficient technique to detect phishing. This approach works efficiently in large dataset. Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99%. An advantage is for the user to make online payments securely. If there is no Internet Connection then the system won't work is the major disadvantage of the project.

II.LITERATURE SURVEY:

This paper provides an overview of features used in their experiments. They presented the data used. In they also described and presented the different experiments conducted and also presented the performance results of various machine learning algorithms. Finally, in last part they provided concluding remarks. There exist a number of different structural features that allow for the detection of phishing emails. In their approach, they make use of sixteen relevant features. The features used in their approach are described. To implement and test their approach, they have used two publicly available datasets i.e., the ham corpora from the Spam Assassin project as legitimate emails and the emails from Phishing Corpus as phishing emails**.** To evaluate the implementation, they used different machine learning methods and a clustering technique on phishing dataset. They used Support Vector Machines (SVM, Biased SVM & Leave One Model Out), Neural Networks, Self Organizing Maps (SOMs) and K-Means on the dataset described. [1].

The overall approach, first described in centers on extracting information that can be used to detect deception targeted at web users, which is accomplished by looking at features from each incoming email or potential attack vector. This process involves extracting data directly present in the email, as well as collecting information from external sources. The combination of internal and external information is then used to create a compact representation called a feature vector, a collection of which are used to train a model. Based on a given feature vector and the trained model, a decision is made as to whether the instance represents a phishing attack or not. They presented a detailed description of the approach, which filters approximately 96% of phishing emails before they ever reach the user. The remainder of this paper is organized in the following manner. First it discusses previous approaches to filtering phishing attacks, while next it gives an overview of machine learning and how we apply it to the task of classifying phishing emails, and how it could be used in a browser toolbar. Thirdly it covers the results of empirical

ISBN : 978-81-943584-3-5

evaluation, as well as some challenges presented therein. [2].

In this paper, they provide a framework for detecting a malicious web page using artificial neural network learning techniques. In addition to the significant detection rate, the objective is to find also which discriminative features characterize the attack and reduce the false positive rate. The algorithm is based on two features group, the URL lexical and the page content features. The experiments has shown the expected results and the high false positive rate which produced by machine learning approaches is reduced.[3].

In this study, an application called "Anti Phishing Simulator" was developed to check the text content and determine whether the related message contained phishing elements. Today, an e-mail can be found in primitive ways whether it is a phishing message or not. For this are looked where this e-mail came from, whether a link with the message matches the actual website, whether the email or referrer web site is using some emotional or exciting words to get a response, whether it is spelling or grammar errors in the email or on the website. However, many people pay attention to this point unconsciously entering the links given to others accounts.[4].

The performance measures of certain controlled machine learning techniques categories such as Bayes algorithms, Bayesian algorithms, tree algorithms, artificial neural network, and support vector machines for classification of a spam email collection held by the UCI Machine Learning Storage have been compared in the study of Panigrahi P. K. The purpose of the work done is to examine the contents of the emails, to learn the limited data set available, and to develop a classification model that can predict whether an email is spam.[5].

They have described how a machine can able to judge the URLs based upon the given feature set. Specifically, they described the feature sets and an approach for classifying the given the feature set for malicious URL detection. When traditional method falls short in detecting the new malicious URLs on its own, they proposed method can be augmented with it and is expected to provide improved results. Here in this work, we proposed the feature set which can able to classify the URLs. The Future work is to fine tuning the machine learning algorithm that will produce the better result by utilizing the given feature set. Adding to that the open question is how they can handle the huge number of URLs whose features set will evolve over time. Certain efforts have to be made in that direction so as to come up with the more robust feature set which can change with respect to the evolving changes.[6].

ISBN : 978-81-943584-3-5

III. PROCESSED SYSTEM:

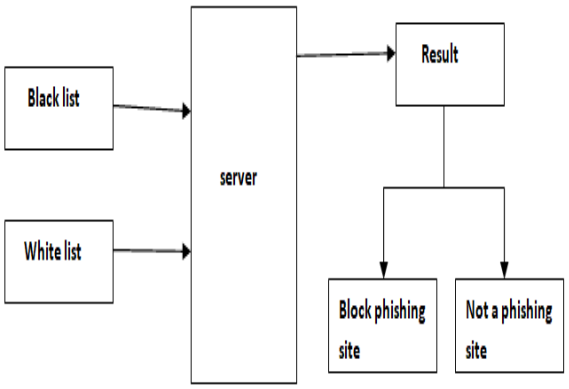

Figure.1: Block Diagram for Detection of Phishing Website

This methodology involves Input the URL to the server shown in figure 1.1. Compare the URL using blacklist and white list approach. Blacklist approach 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 database and gives warning. This approach is easy and faster to implement as it see URL is in database or not. The White list approach is used as filter that blocks Phish WebPages used to imitate innocuous user behavior. If the URL is white listed then feature extraction using Visual Similarity Approach takes place else the phishing will be blocked.At the end predicting the result takes place to check whether it is aphishing site or not.If YES then Block the site or else it is not a phishing website.Phishing website can be tracked through the methods like location based login. In location based login when the user login into webpages it will check the distance and time of the login.In this case if the login time of the user is done before the accurate time then the account should be blocked.The Another method is to track the IP address of the webpages like facebook,gmail through whitelist and blacklist.The last method is detecting the exact pages by matching the original facebook page or any webpages with other dummy pages.If the original page matches with the other dummy pages then it is said to be the Phishing webpage.

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 ofphishing.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 thanother.

ISBN : 978-81-943584-3-5

REFERENCES:

[1] *"Detection of Phishing Attacks: A Machine Learning Approach*" Ram Basnet, Srinivas Mukkamala, and Andrew H. Sung, New Mexico Tech, New Mexico 87801, USA.

[2] *"Learning to Detect Phishing Emails"* Ian Fette, Norman Sadeh, Anthony Tomasic, School of Computer Science Carnegie Mellon University Pittsburgh, PA, 15213.

[3] Malicious Web Page Detection: A Machine Learning Approach Abubakr Sirageldin, Baharum B. Baharudin, and Low Tang Jung Computer & Information Science Department, University Technology Pertonas Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia.

[4] *"Detection of phishing attacks"* Muhammet Baykara , Zahit Ziya Gürel, Department of Software Engineering, Faculty of Technology Firat University, Elazig, P. K. Panigrahi.

[5] *"A Comparative Study of Supervised Machine Learning Techniques for Spam E-mail Filtering,"* 2012 Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, pp. 506-512, 2012.

[6] *"Detection of Malicious URLs using Machine Learning Techniques"* Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-4S2 March, 2019.