Original Research Article

J. Computing & Int. Systems (2021) 178–181

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

SACRED HEART RESEARCH PUBLICATIONS

ISSN: 2456-9496

Features Classification using Learning Machine for website phishing

Ajit V. Gaikwad[#1], **Pradip M. Jawandhia**[#2], **Sachin Manohar Dandage**[#3]

Received on 07th JUN 2021, Accepted on 21st JUN 2021

Abstract — To perform malicious misconduct and perform various fake transactions by stealing the information on an individual makes phishing the most vulnerable attacks. The websites had various links into their content during browsing. The main of the study is to identify the phishing website based on some classification patterns and helps to generate the better result against phishing. Various algorithms and methods are already implemented to rectify the phishing website whereas SVM, Naives Bayes provides better result among all methods

Keywords - Extreme Learning Machine, Features Classification, Information Security, Phishing.

I. INTRODUCTION

Many works were handled on the basis pf internet and by virtue of which it becomes one of the most important tool to share vulnerable information. And its very increasing popularity in terms of use makes it the most powerful interface for various interactions as well as transactions. Precautionary measures are the main intension to provide the security to all this transaction make it possible to maintain the security essentials for the same. Even though the major role in response to the security point of view is to secured the information from various threats and links which make the user to unevenly make use of the resources without any notifications and makes it easy to alter and steal the information. This information carries the links on which the users directly get access to enter the username password id and other various important information and lost all the credentials to it. Phishing websites contain various hints among their contents and web browser-based information. It always intends to fetch the vulnerable information from the user through a valid resource link or any transaction link. This links looks so authentic that user unintentionally drag to this links and put all the data into it, which became a prone to the leakage of data and information.

II. WEBSITE PHISING

Properties defined through email point of view makes the pattern of phishing website through but always tries to search the victim who are eagerly waiting for such messages through various website and other pop-up links that direct them to the malicious website. On the resemblance with that we introduced the method which make the use of some neural network and Naives Bayes algorithm to implement the system which will help to detect the phishing website with some statistical and language processing database. The method of classification uses certain approach which to some extent has less efficiency and scalability. They make the use of structural properties and some content properties which helps to reduce the errors during analyzing the website. The method of link guard was also introducing that basically works on detecting the malicious hyperlinks which directs to the phishing websites. Comparisons of DNS with actual website address that can be able to checks the decimal dots in IP address which also has some fake IP address. Since it helps to detect the fake IP address but the main drawbacks that even though sometimes this fake IP has correct address but the link and pop up generated through it may have the phishing website links so sometimes it may fall into false detection of IP addresses. The phish was the generous concept design by MC Grath and Gupta but as they referring the open directory for searching of phishing website it may create the fake phish tank by storing the non-phishing website into it which creates mixing of both links and creates the problem for the user to identify form the original websites. This link contains many vulnerable information such as original id's, URL detailed information about width, length and character distribution parameters.

* Corresponding author: E-mail: [1]ajit.gaikwad007@gmail.com, [2]plitprincipal@gmail.com, [3]dandage.sachin@gmail.com

[1]Principal Pankaj Laddhad Institute of Technology and Management Studies Buldhana,

[2]Head of Department Computer Science Pankaj Laddhad Institute of Technology and Management Studies Buldhana

[3]Pankaj Laddhad Institute of Technology and Management Studies Buldhana, Sant. Gadgebaba Amravati University Maharashtra India

.

2.1 BLACLIST AND WHITELIST APPROACH

Whenever an individual visited a blacklist or whitelist website it's always necessary to identify whether it's a phishing or legitimate website. Newly created website was not identified through blacklist and whitelist website which ultimate proceed further to phishing website links and it's the basic disadvantages of whitelist and blacklist website.

2.2 INTELLEGENT HEURISTICS APPROACH

This approaches mainly works on the principle of gathering and evaluating the most influential and important information from the website and make it possible to detect the legitimate website from phishing website. This data can be used to generate the dataset for many detection platforms. And then Machine Learning helps to classify the difference between legitimate ad phishing websites by training this dataset. When the training was done the classifier works efficiently to identify the website in terms of phishing and legitimate website which was hidden during training phase. Hence this approach is very efficient to detect the newly created website which was impossible to rectify in phish tank methods.

2.3 MACHINELEARNING

It always focuses on managing and developing various computational flows of sequences and generates rules focuses on generating number of instances to generate a unique model to predict the website as phishing and legitimate one. When the instances are labeled with known facts we proposed the method as supervised learning and when it's not labeled it was proposed as unsupervised learning. There are many machine learning techniques available such as support vector machine, naïve bayes classifier, back propagation neural network, random forest, k-nearest neighbor algorithm and many which helps to design the dataset and categories the rules in accordance to predict some rules to identify the features of any dataset

2.4 URL FEATURE ANALYSIS

The identification of phishing website can also be determined from URL identification to understand whether it's a phishing website or legitimate website. The tag attributes help to identify the overlap values which consist of summation of selected attribute values in combination with other attributes.

2.4.1 FINDING ATTRIBUE VALUES?

The value ranging from -1 to 1 computed to find out the attribute values. Whereas some prefix and suffix also interlinked which each other to identify the URL tag of attribute ranging from the defining values between -1 and 1.

2.5 EXTREME LEARNING MACHINE (ELM)

It is a feed forward artificial neural network model which has only one single hidden layer. For proper evaluation of ANN it is always necessary to make use of threshold value, its weight and its way pf activation that are required to undergo the process of modeling trained data. Input weights in ELM are randomly selected with parameter as gradient based and its output weights are analytically obtained. To obtained the cells associated with hidden layer a linear function, on-linear functions and non-derivable functions are used. JAVA Scripting was also used to visualized the website as the legitimate one since it's the fake appearances of that phishing website, which can be done by placing the images of legitimate website on phishing web addresses. The best example was carried out in the year 2006 against PayPal where its logo image was use to represent the phishing website as phishing website. Flashing based website always sometime used to make the phishing website looks like legitimate website which undoubtedly attracts the user to access and share its vulnerable information.

III. SOME SYSTEM USED FOR PHISHING DETECTION:

ANOMALY BASED PHISHING DETECTION SYSTEM:

An identity of any website is unique and whenever the phisher tries to copy such website it always keep some identity on the phishing website. This in turns helps the website detection to check such anomalies and distinguished that website among legitimate website and phishing website using some object classifier and some transactional history. Two identity extractor and page classifier was defined to find out the anomalies between various websites.

3.1.1 THE IDENTITY EXTRACTOR:

The keyword extraction algorithm in the information retrieval is used for Identity extraction. The webpage identity is determined by DOM objects in this method as follows: -

I. Title: the title of one web page (namely, the text between the tag <title> and < \title>).

II. Description: the content property of the META whose name or http-equiv is "description".

III. Copyright: the content property of the META whose name or http–equiv is "copyright".

IV. ALT/title: the alt and title properties of the DOM objects such as IMG, AREA, INPUT, APPLET, OBJECT.

V. Body: the text in the main body or the images in a web page. There are many technologies to recognize texts from one image, such as Optical Character Recognition (OCR).

3.1.2 PAGE CLASSIFIER:

The next method of classifier used was http transaction part. Support Vector Machine for VAPNIK's connects to the webpageand trying to the identity of the current page. And converting the information into vectors and sending to SVM page classifier and classifies the webpage as legitimate or phishing one. Also it extracts the structural feature of the current page through its flow from top to bottom.

3.2 PHONEY: MIMICKING USER RESPONSE TO DETECT PHISHING ATTACKS:

This system provides fake response to the user and creates a fake communication between client and server to access some important incoming mails.

Its working is as follows:

First the incoming message was initiated to the server since identifies as legitimate user it parses into the message body and embedded some links and HTML forms. Then the mail which contains some form link considered as the malicious mail. The once it is identified the control move forward to content scanner and the forms embedded with it was broken down into various elements and with similar associated text. Then this element is compared with Hah Data base. During this period the fake data was filled in the form and this data was used as malicious information to extract various features included into the mail.

3.3 DEFENDING THE WEAKEST LINK: PHISHING WEBSITES DETECTION BY ANALYSING USER BEHAVIOUR

The User Behavior Based Phishing Detection analyzes the online behavior of the user while sending some vulnerable credentials to the website or saving it onto the already logged system.

It basically contains three modules:

I. User Profile creation in which many basic data was taken from the user which makes the website in white listing.

II. It always monitors the data sending of the user and start detection engine to identify the user profile credential.

III. The detection engine when activates works on two basic modes one is training mode and second is detection mode. Profile update was done in training mode whereas in detection mode it checks whether the website is phishing website or legitimate. If it's a legitimate one, then that was not a problem but if it's a phishing website it sends alert to user behavior based phishing detection asking for the credentials one by one related to every website which was accessed accordingly. After the process of detection of website, it activates the Train user behavior based phishing detection and binds the website to the blacklist if it has some phishing type resemblance. And this process will continue until the user behavior method helps to find out the list of black listed website, this approached was listed by ALEXA in network traffic accessibility.

IV. PROPOSEDWORK

The work starts by creating the database, which was classified by considering the input parameters using the learning mechanism. Support Vector Machine and Naives Bayes classifier are used to identify the difference between phishing website and authentic website. The database which was created was created in such a way that it makes identification of phishing website perfectly and providing maximum accuracy than the previous methods. Building a model with such perfection needs the dataset to be precisely refined so as to generalize good performance and result.

4.1 ALGORITHM:

An algorithm mentioned below includes the way to detect the website whether it's real or phishing website through some parameter (categorizing the input URL into words, letters and phrases as a part of feature extraction) classification has been done. The wordings present in the website extracted the feature in terms of words, phrases and letters and verify with the database mention

1. During the process the dataset was prepared which includes the words and phrases set
2. Once the pre process completes the features will be extracted from the URL
3. Compute attribute values, if
   a. Attribute present value = 1
   b. Attribute absent value = -1
   c. Attribute not considered = 0

   3.1 Select attribute X and Y

   3.2 Compute equation for X and Y
4. Calculating the threshold with exact matching of the dataset
5. Finding the range value.
6. Distinguish phishing and legitimate site using attribute value and comparing the dataset.
7. It also helps to compute Sensitivity and Specificity. [4][6]

V. ADVANTAGES

- Build an authenticate communication bridge between transferring agent and mail user agent.
- Classifier on phishing mail detection was clearly understandable.
- Accuracy is high as compared to previous method.
- Process of classification is good.

VI. DISADVANTAGES

- Time consuming
- Huge number of features
- Consuming memory

VII. CONCLUSION

Many automatic information processing applications are used to avoid the manual data entry process through many websites. The data which we entered into the system generated the processed information as an output. Due to extensive used of website in many fields such as business, trading, financial transaction, scientific data cloud storage, secured listing cloud information and many more vulnerable websites. The way we interact with the website makes our data openly available to such hackers and malicious and phishing website. The phishing attacks became the most dangerous type of attacks in its terms which can steal the vulnerable information of your user profile even though of your bank transactions details. To avoid such attacks a classification model was defined with various parameters and feature included in it to stop the threats at highest accuracy.

REFERENCES

[1] G. Canbek and Sarolu, "A Review on Information, Information Security and Security Processes," Politek. Derg., vol. 9, no. 3, pp. 165– 174, 2006.International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8S2, June 2019

[2] Vol-4 Issue-6 2018 IJARIIE-ISSN(O)-2395-4396 9322

[3] Vol-4 Issue-6 2018 IJARIIE-ISSN(O)-2395-4396

[4] L. McCluskey, F. Thabtah, and R. M. Mohammad, "Intelligent rule based phishing websites classification," IET Inf. Secur., vol. 8, no. 3, pp. 153–160, 2014. R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Comput. Appl., vol. 25, no. 2, pp. 443–458, 2014.

[5] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," Internet Technol. ..., pp. 492–497, 2012.

[6] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," Appl. Soft Comput. J., vol. 48, pp. 729–734, 2016.

[7] N. Abdel hamid, "Multi-label rules for phishing classification," Appl. Comput. Informatics, vol. 11, no. 1, pp. 29–46, 2015.

[8] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machinelearning-based web phishing.