IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detection Of Phishing Websites Using Data Mining

[1]Mr Aniket Kote, [2]Mr Sanket Kharche, [3]Mr Pravin Aware, [4]Mr Abhishek pangavhane

[1]Department of Computer Engineering,
[1]SRES COE, Kopargaon, India

Abstract: Recently many Cyber attacks have increased because of growing use of internet. Among all this the well-known cyber threat we find increasing day by day is phishing attacks. Phishing is where the victim's credentials are obtained by an illegitimate website. The aim of these phishing websites is to acquire confidential information about username, password and banking credentials of the victim. Phishing websites look similar to the illegitimate website so the user can't differentiate among them. This paper proposes a system which will detect old as well as newly generated phishing URLs that have completely no past behaviors to judge upon, using Data Mining. The main objective of this system is to develop a Chrome browser plugin that detects phishing sites in real-time while the user browses the page. The model will be trained with exhaustive dataset so that we can assure maximum accuracy.

Index Terms **-Phishing, Data Mining, Web Security, Chrome Extension, Random Forest Classifier.**

1. INTRODUCTION

Phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords, and credit card details (and money), often for malicious reason. It is typically carried out by email spoofing or instant messaging, and it often directs users to enter personal information at a fake website, the look and feel of which are identical to the legitimate site, the only difference being the URL of the website in concern. Communications purporting to be from social web sites, auction sites, banks, online payment processors are often used to lure victims. Detecting phishing websites often include lookup in a directory of malicious sites. Since most of the phishing websites are short lived, the directory cannot always keep track of all, including new phishing websites. So, the problem of detecting phishing websites can be solved in a better way by different techniques. Based on a comparison of different techniques, the random forest classifier seems to perform better. Only way for an end user to benefit from this is to implement detection in a browser plugin. So that the user can be warned in real time as he browses a phishing site.

2. MOTIVATION

. As we already have different techniques to identify this phishing website. Most popular one of this kind is Phish Tank. According to Phish Tank, it is a collaborative clearing house for data and information about phishing on the Internet. Also, Phish Tank provides an open API for developers and researchers to integrate anti-phishing data into their applications at no charge. Also, Google has an API called Google Safe Browsing API which also follows directory-based approach and also provides open API similar to phish Tank. But this kind of approach clearly can't be effective as new phishing web sites are continuously developed and the directory can't be kept up to date always. This also leaks users browsing behavior as the URLs are sent to the phish Tank API. So, this System not only detect the phishing website but also aims to implement the same in browser plugin removing the need of external web service and improving user privacy.

3. LITERATURE SURVEY

Cyber Threats is the major issue increasing all over the world because of increasing use of internet. Many different approaches are available as many people tried many solutions. Before designing this system we surveyed many papers fallowing is their overview:[1] First paper we studied was "Real time detection of phishing websites" by Abdulghani Ali Ahmed, Nurul Amirah Abdullah published in 2016.The main drawback of that system was that the accuracy of this heuristic-based depended on the discriminative features that may help in distinguishing the type of website whether it is a legitimate or phishing site, hence it was not that continent to be used in real time.[2] Second paper we studied was "Comparative analysis of features based machine learning approaches for phishing detection" by Ankit Kumar Jain proposed in 2016.Altough this was the better approach than the previous one of directly studying the features instead of whole URL but the time required and the regular update and easy use was the main issue in this system.[3] The next paper we studied was "A Novel Algorithm to Detect Phishing URLs" by Varsharani Ramdas Hawanna, V.Y Kulkarni proposed in 2016.The problem with this system was it was effective only for HTTP URLs. The last paper we studied was "A Hybrid Model to Detect Phishing- Sites Using Supervised Learning Algorithms" by Tahir, M. Amaad Ul Haq proposed in 2016, But the drawback with this system was the algorithm complexity and it was not possible to be used on large scale.

4. PROPOSED SYSTEM

In the proposed system, a web host model is used for phishing website detection. The model will be based on classification algorithm and will be trained using a training dataset. This model will be deployed online, which will directly communicate with the chrome extension. The detection of the phishing website will be based on URL and website attributes. This system will be an integration of all the functions carried out on the client and server side. On the server side, there will be a classifier model trained using the random forest algorithm; whereas on the client side, a chrome extension will be built and added to the chrome web browser. When the client visits any website using the chrome browser, the URL for the same can be fetched by the chrome extension. From the URL and the displayed webpage, various attributes of the website can be extracted. These extracted URL attributes will act as test data for the classifier deployed in the cloud which can be trained on a phishing website dataset. The classifier will predict the entered URL as either malicious or safe. If it is a phishing website then the user will be alerted that, if they proceed further on this URL their credentials are at risk of getting misused and if it is a safe website then the user can carry on further operations on that page.

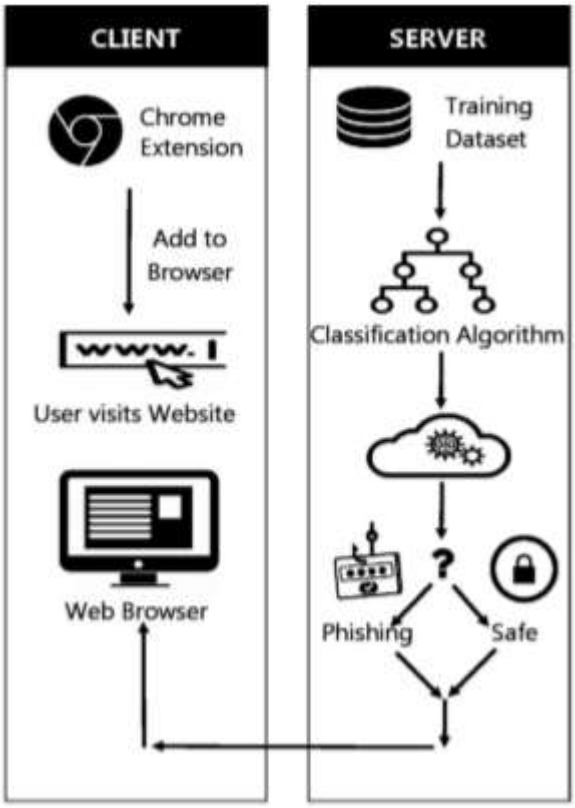

Fig.system architecture

4.1 Chrome Extension

They are small software programs that customize the browsing experience. They enable users to tailor **chrome** functionality and behavior to individual needs or preferences. They can range from a simple icon, such as the Google Mail Checker **extension** shown on the right, to overriding an entire page. We are creating plugin and that plugin is added to the chrome extension.

4.2 Pre-processing

The dataset is downloaded from UCI repository and loaded into a numpy array. The dataset consists of 30 features, which needs to be reduced so that they can be extracted on the browser. Each feature is experimented on the browser so that it will be feasible to extract it without using any external web service or third party. Based on the experiments, 17 features have been chosen out of 30 without much loss in the accuracy on the test data. More number of features increases the accuracy and on the other hand, reduces the ability to detect rapidly considering the feature extraction time. Thus, a subset of features is chosen in a way that the tradeoff is balanced. Then the dataset is split into training and testing set with 30% for testing. Both the training and testing data are saved to disk.

IP address	Degree of subdomain	Anchor tag href domains
URL length	HTTPS	Script & link tag domains
URL shortener	Favicon domain	Empty server form handler
@' in URL	TCP Port	Use of mailto
Redirection with '//'	HTTPS in domain name	Use of iFrame
-' in domain	Cross domain requests	

Fig. web features

4.3 Training data

For the proposed model, a publicly available dataset offered by UCI repository will be used for training. It comprises of 11055 records, out of which 4,898 are phishing websites while 6,157 are legitimate websites. And after preprocessing 70% data is used for training using random forest classifier.

4.4 Exporting module

Every machine learning algorithm learns its parameter values during the training phase. In Random Forest, each decision tree is an independent learner and each decision tree learns node threshold values and the leaf nodes learn class probabilities. Thus, a format needs to be devised to represent the Random Forest in JSON.

5.ALGORITHM

We are mainly going to use random forest classifier for the proposed system for training data.

5.1 Random forest algorithm

Random forest algorithm is a supervised classification and regression algorithm. As the name suggests, this algorithm randomly creates a forest with several trees. Generally, the more trees in the forest the more robust the forest looks like. Similarly, in the random forest classifier, the higher the number of trees in the forest, greater is the accuracy of the results.

Step 1: First, start with the selection of random samples from a given dataset.

Step 2: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3: In this step, voting will be performed for every predicted result.

Step 4: At last, select the most voted prediction result as the final prediction result.

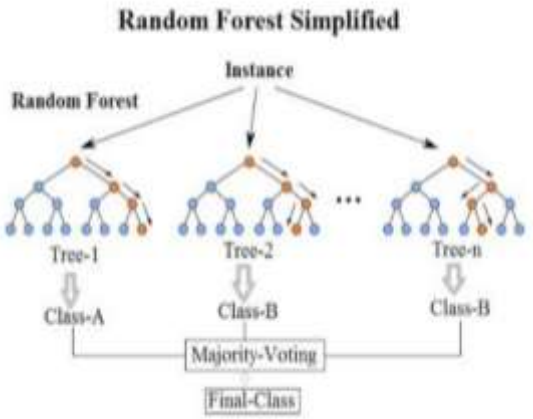

Fig Random Forest Classifier

5.2 Why Random Forest Algorithm

- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance then single decision tree.
- Random forests are very flexible and possess very high accuracy.
- Random Forest algorithm maintains good accuracy even a large proportion of the data is missing.

6.CONCLUSION

The Proposed System's aim is to implement the detection of the phishing websites using data mining. This task will be done by extracting the features of the website via URL when the user visits it. The obtained features will act as test data for the model. Random Forest Algorithm can be used to train the proposed model. The main task of this system is to detect the phishing website and alert the user beforehand so as to prevent the users from getting their credentials misused. If any user still wishes to proceed, it can be done at their own risk.

REFERENCES

[1] Ahmed, Abdulghani Ali, and Nurul Amirah Abdullah."Real time detection of phishing websites." Information Technology, Electronics and Mobile Communication Conference (IEMCON), 7th Annual. IEEE, 2016.

[2] Jain, Ankit Kumar, and B. B. Gupta." Comparative analysis of features-based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on.IEEE, 2016,

[3] Hawanna, Varsharani Ramdas, V. Y. Kulkarni, and R. A. Rane." A novel algorithm to detect phishing URLs." Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016.

[4] Tahir, M. Amaad Ul Haq, et al." A Hybrid Model to Detect Phishing- Sites Using Supervised Learning Algorithms." Computational Science and Computational Intelligence (CSCI), 2016 International Conference on. IEEE, 2016.