Phishing Website Detection using ML

Abhijith A
Dept. of Computer Science
TKM Institute of Technology
Karuvelil Kollam Kerala

Mishal K K
Dept. of Computer Science
TKM Institute of Technology
Karuvelil Kollam Kerala

Beulah Susan Koshy
Dept. of Computer Science
TKM Institute of Technology
Karuvelil Kollam Kerala

Sapna Shaji S
Dept. of Computer Science
TKM Institute of Technology
Karuvelil Kollam Kerala

Shemimol B
Assistant Professor,CSE
TKM Institute of Technology
Karuvelil Kollam Kerala

***Abstract* - The purpose of this project is to design an intelligent system for detecting phishing websites. Phishing is one of the social attack which aims in stealing sensitive information of the users such as login credentials, credit card numbers etc. Here we have collected phishing dataset from phish Tanks as well as from phishing sites and are compared with the algorithms which classifies the phishing dataset into phishing or legitimate. We propose a web application for detection. The algorithm used is random forest in order to get better performance and accuracy. This system uses a database in order to store phishing websites which are already tested and can be used as blacklist, which makes the classification even faster, as it reduces repetition.**

Keywords - Phishing site; Random Forest algorithm; URL

I. INTRODUCTION

Phishing (Fig.1) is a cyber-attack. The goal is to retrieve personal information of the recipient by making them believe that the message is something they want or need i.e., request from the bank etc. Phishing employs two phishing techniques

malware and deceptive based phishing. The attack can occur in two ways either by receiving suspicious email that led to fraud site or by users accessing links that go directly to a phishing website.

In general, two approaches are employed in identifying the phishing sites. The first one is based on blacklist. It works by comparing the requested URL with those in that list but this approach has a drawback that is it cannot identify a new fraud site which has been created within a fraction of second. The latter approach is heuristic based approach. In this approach

, several features are collected from various website to classify it into either phishing or legitimate.

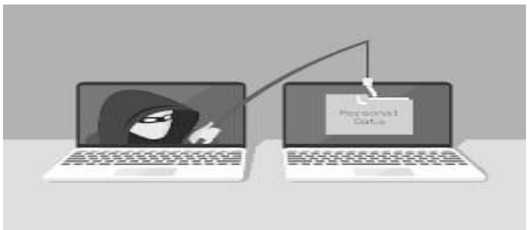


Fig.1. Phishing

Our paper resolves all the drawbacks of the existing system. This project employs machine learning technique for predic- tion task and supervised learning algorithm namely random forest technique for exploring results. Supervised learning is a type of machine learning technique in which labelled training data are used to train the machine and on the basis of that data it predicts the output. Random forest is used for classification and is a machine learning algorithm that belongs to supervised learning technique. For the successful detection of phishing site, it should detect in less time and with high accuracy.

Prediction and prevention of phishing attack is very crucial step towards safeguarding online transaction. The aim is to develop a model to safeguard users from phishing attacks. This can be done using unique features of phishing websites. The goal of our paper is to develop a web application which notifies the user when it detects a phishing site.

The rest of the paper is organized as follows. Section2 reviews the background study on the detection of phishing sites using various approaches. Section 3 explains methodology and working of our application. Finally, we conclude the paper by giving the key points of our entire work in section 4.

II. BACKGROUND STUDY

A. *Detection Using Heuristic Approach*

Srinivasa Rao and Syed taqi has introduced a desktop application system [1] called Phish shield to detect the phishing webpages. It concentrates on URL and website content of phishing page. The input to this system is URL and it outputs the status of URL as either suspicious or trusted website. Here heuristic approach is employed which studies the structure of content and URL of phishing websites to extract the features of the phishing sites. It then designs the model to detect phishing site based on the extracted features. Website identity, title content, copy right content, zero links in the body of the HTML are the heuristic approaches to detect phishing. Phish shield is able to detect zero hour phishing attack. The tool used to develop phish shield application include NetBeans
8.0.2 IDE, JAVA compiler, JSoup api and firebug. Even if the content can be replaced with images, this application is also able to detect phishing sites.

B. *Case Based Reasoning Technique*

Hassan and Abdelfettah introduced Case Based Reasoning(CBR) Phishing Detection System[3].The core part of the system is CBR methodology. To predict a solution for a problem, it uses historical data namely experiences or cases. The main feature of this system is that it works similar to human thinking such that it can solve problems from past experiences as similar problem have similar solution. It consist of four phases; Retrieve, Reuse, Revise and Retain. This technique is designed to be updated with approved phishing attack experiences. There are two types of experiences such as offline and online experiences. It can detect new phishing attack even if the dataset is small. So the system is highly dynamic and adaptive.

C. *Content Based Approach*

Yue Zhang and Jason proposed content based approach for detecting phishing website[2]. Cantina make use of TF-IDF information retrieval algorithm. It is used for comparing and classifying documents as well as retrieving documents from a large corpus. Robust hyperlink is an application of TF-IDF and the basic idea is to provide independent descriptions of networked resources that is URL. Cantina works as follows. When we are given webpage, it calculates the TF-IDF scores of the webpage based on each term. It takes five terms with highest TF-IDF weights and generate a lexical signature. In this case, we use Google as the search engine to feed this lexical signature. If the domain name of N top search result matches with the domain name of the given webpage, Then it considers website to be a legitimate one, otherwise it is considered as the phishing website. Cantina is implemented as the Microsoft Internet Explorer extension.

D. *Machine Learning Approach*

Sirageldin, B. B. Baharudin, and L. T. Jung, proposed a framework for detecting malicious webpage using Artificial Neural Network[5]. The algorithm is based on two group of features namely URL lexical and page content features. Feature extractor, Learning and model content features. Fea- ture extractor, learning and model selector and Detector are the three components of the model. Webpage features are extracted by feature extractor, Here we collects the dataset. Learning algorithm with best result is selected in learning and model selector to build the model. Final classification model generated from the learning and model selector component by using detector.

III. PROPOSED SYSTEM

A. *Methodology*

In this system, the user will login using his/her user ID and password. After successful authentication, the user will be di- rected to the home page where the user can enter the suspicious URL which needs to be tested or can view all the phishing URLs already existing in the database. After entering the URL, the first step is to check whether the database(blacklist) already have entered URL marked as phishing. If found, the available result or information of the corresponding URL is obtained from the database and presented before the user in the GUI of the user. Else, the system takes the URL to be tested by using machine learning technique. It will take the URL and classify using Random forest technique, which is a popular machine learning technique used for classification processes. Random forest technique is also proven to be the best machine learning technique which provides with the most precise result. After the classification, if the URL turns out to be that of a phishing site, the resultant information will be published in the GUI and also can be pushed or uploaded to the database(blacklist) with the users permission. The architecture is shown in fig 2.

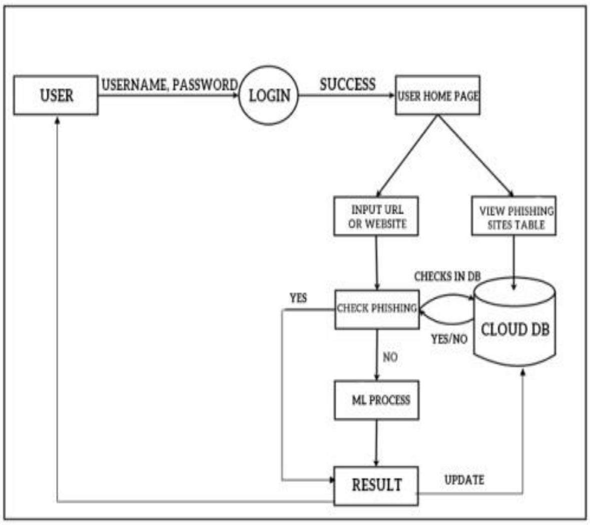


Fig. 2. System Architecture

1) Data Collection: The datasets were collected from Phish tank archieve and google search operators. It consist of 2456 instances and 18 features. Attribute values are in the form of integer 0 and 1 where 0 represent legitimate and 1 represent phishing. First the dataset has been processed to get matured data in desired format, Then it is divided into two section as training 70% and testing 30%. The experiment have been carried out using our python of phishing website based application. Random forest algorithm have been used in the work for detection of phishing website.

2) Random Forest: The Random Forest algorithm has char- acteristics such as high robustness and performance. We use machine learning architecture to predict the phishing sites. The dataset contains many data with information of phishing or not. The datas are taken and passed to the algorithm. The random forest classifier is used for classification. The random forest algorithm is made up of thousands of decision trees and then gets the prediction from each of them and at last, by means of voting, best solution is selected.

Steps involved in this system are:
1. Feature Extraction
2. Word to Vector Conversion
3. Training the model using Random forest Algorithm 4.Testing the model by inserting a URL

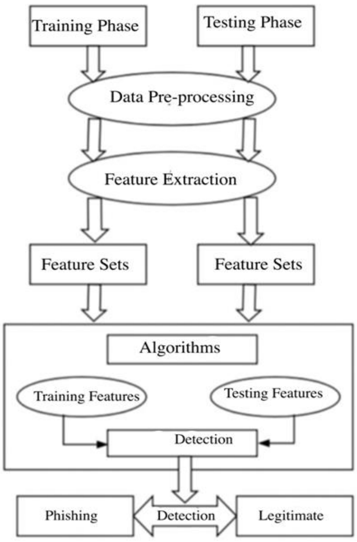


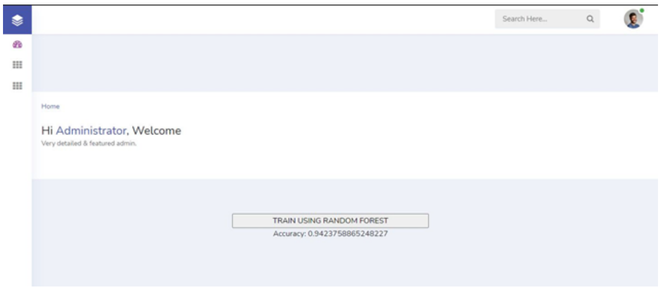
Fig.3.Phases of a system

B. Advantages

- This technique provides the phishing webpage detection by hiding the users identity from phishers.
- The computation time is very less.
- It can detect harming attacks, which are undetectable by many existing system.

IV. RESULT AND DISCUSSION

First the Users and admins log on to our system by login page. From here the admins approves new user and trains the system.

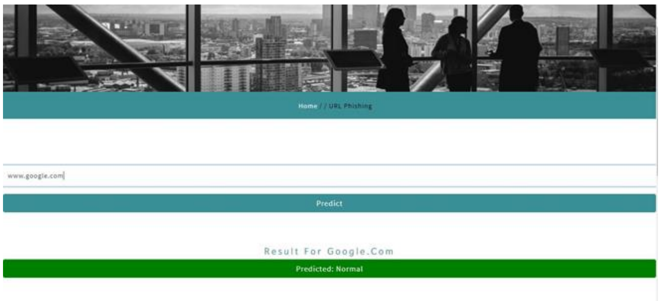


Hi Administrator, Welcome
Very detailed & featured admin.

Accuracy: 0.9423758865248227

fig 4: Admin page

www.google.com is not a phishing website, so the output predicted is “Normal”.

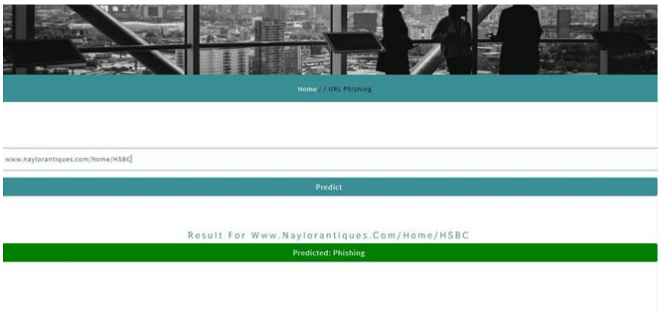


www.google.com

Result For Google.Com
Predicted: Normal

fig 5 Non phishing website

www.Naylorantiques.com/Home/HBSC Is a phishing website, so output predicted is “Phishing”.



www.naylorantiques.com/home/HSBC

Result For Www.Naylorantiques.Com/Home/HSBC

Predicted: Phishing

fig 6 phishing website

V. CONCLUSION AND FUTURE SCOPE

Phishing is a growing problem for internet users. There are a number of anti-phishing tools available to cope against this problem. Still there are limitation on accuracy because detec- tion techniques are time consuming. Among several machine learning algorithm, Random-forest gives the better result. This work become unique from other existing work by proposing a group of features that can be extracted automatically using our own software tool. In future we can make the system available in mobile devices.

REFERENCES

[1] R. S. Rao and S. T. Ali, ”PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.

[2] Y. Zhang, J. I. Hong, and L. F. Cranor, ”Cantina:A Content-based Approach to Detecting Phishing Web Sites,” New York, NY, USA, 2007,pp. 639-648.

[3] Hassan Y.A.Abutair, AbdelfettahBelghith, “Using Case-Based Reason- ing for Phishing Detection,” in 2017 .The 8th International Conference on Ambient SystemsNetworks and Technologies(ANT) 2017,pp.281- 288.

[4] L. Breiman, ”Random Forests,” Machine Learning, vol.45, no. 1, pp.5-32, Oct. 2001.

[5] Sirageldin, B. B. Baharudin, and L. T. Jung, ”Malicious Web Page Detection: A Machine Learning Approach,” in Advances in Computer Science and its Applications, Springer, Berlin, Heidelberg, 2014, pp.217- 22