Real-Time Phishing Detector Browser Extension - PHISHERMAN

Prof. Prashant S. Gawande[1], Mr. Tanmay R. Shrimali[2], Mr. Bhushan N. Dhole[3], Mr. Tanmay S. Ahire[4], Mr. Pawan P. Patil[5]

[1] *Professor, Department Of Information Technology, Sandip Polytechnic, Nashik, Maharashtra, India*
[2] *Student, Department Of Information Technology, Sandip Polytechnic, Nashik, Maharashtra, India*
[3] *Student, Department Of Information Technology, Sandip Polytechnic, Nashik, Maharashtra, India*
[4] *Student, Department Of Information Technology, Sandip Polytechnic, Nashik, Maharashtra, India*
[5] *Student, Department Of Information Technology, Sandip Polytechnic, Nashik, Maharashtra, India*

ABSTRACT

Phishing is defined as imitative a worthy company's website intent to take private information of a person. In order to eliminate phishing, different methods planned. However, only one single remedy cannot eliminate this danger completely. Data mining is a likely technique used to discover phishing attacks. In this paper, an clever system to detect phishing attacks is given. We used various data mining methods to resolve group of websites: authorized or phishing. Various classifiers were used in order to build faithful intelligent system for phishing website detection. Classification quality, area under receiver operative characteristic (ROC) curves (AUC) and F-measure is used to measure the operation of the data mining techniques. Results showed that Random Forest has carry out best among the classification methods by achieving the higher accuracy 97.36%. Random forest run-times are quite fast, and it can deal with various websites for phishing detection.

Keyword : - *Data Mining, Phishing Websites, Websites Threats, Machine Learning, Cyber Security*

1. INTRODUCTION

This is a phishing website detection program that focuses on client side implementation with rapid & real-time detection so that the users will be warned before getting phished. The main implementation is porting of Random Forest classifier to JavaScript. Similar works often use web-page features that are not feasible to extract on the client side and this results in the detection being dependent on the network. On the other side, this system uses only features that are possible to extract on the client side and thus it is able to provide fast detection and better privacy it increases the usability of the system. This work has identified a subset of web-page feature that can be implemented on the client side without much effect in accuracy of results.

2. PHISHING DETECTOR

This is a phishing detection plugin for browser that can detect and warn the user about phishing web sites in real-time using random forest classifier. Intelligent phishing website detection using machine learning, the random forest classifier seems to outperform other techniques in detecting phishing websites. One common approach is to make the classification in client side and then let the plugin shows for result. This project aims to run the classification in the browser itself. The advantage of classifying in the client side browser has advantages like, better privacy (the user's browsing data need not leave his machine), detection is independent of network latency. This project is mainly of implementing machine learning techniques in JavaScript for it to run as a browser plugin. Since JavaScript doesn't have much ML libraries support and considering the processing power of the client machines, the

approach needs to be made lightweight. The random forest classifier needs to be trained on the phishing websites data-set using python scikit-learn and then the learned model parameters need to be exported into a portable format for using in JavaScript.

3. OBJECTIVES

1. To implement phishing detection techniques in browser via extension or plugin.
2. Better privacy user browsing data and patterns should not be leaked. The user browsing history never leaves the browser.
3. Network independent the detection should not be affected by network latency.
4. Rapid detection with less computing power.
5. More accuracy & security in web surfing.

4. SYSTEM ARCHITECTURE

A Random Forest classifier is trained on phishing sites data-set using python scikit-learn. A JSON format to stand for the random forest classifier has been devised and the learned classifier is exported to the same. A browser script has been implemented which uses the exported model JSON to separate the website being loaded in the active browser tab. The system intent at informing the user in the event of phishing. Random Forest classifier on 17 features of a website is used to classify whether the site is phishing or authorized. Features are selected on basis that they can be extracted entirely offline on the client side without being dependent on a web service or third party. The data-set with chosen features are then isolated for training and testing. Then the Random Forest is trained on the training data and exported to the above mentioned JSON format. The JSON file is hosted on a URL. The client side chrome plugin is made to execute a script on each page load and it starts to extract and encode the above chosen features. Once the features are encoded, the plugin then checks for the exported model JSON in cache and downloads it again incase it is not there in cache. With the encoded feature vector and model JSON, the script can run the classification. Then a warning is shown to the user, in-case the website is classified as phishing. The entire system is designed lightweight so that the detection will be speedy.

Table -1: SWOT Analysis

STRENGTHS	WEAKNESSES	OPPORTUNITIES	THREATS
1) Enables user privacy. 2) Rapid detection of phishing. 3) Can detect new phishing sites too. 4) Can interrupt the user in-case of phishing.	1) JavaScript limits functionality. 2) Cannot use features that needs a external service such as SSL, DNS, page ranks. 3) No library support.	1) Everyone awake of privacy and security can use this plugin. 2) Non technical persons who do business transactions are vulnerable to phishing and they are potential end users for this.	1) Server side classification plugins may execute better than this and users without privacy concerns may opt of those. 2) Browser Plugin API will be endlessly changed.

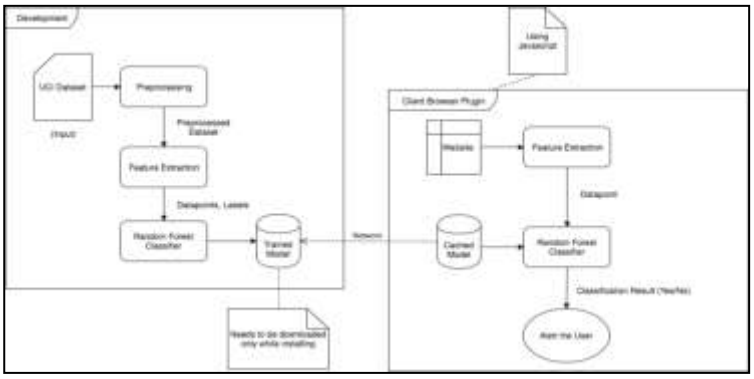

Fig-1: Block Diagram

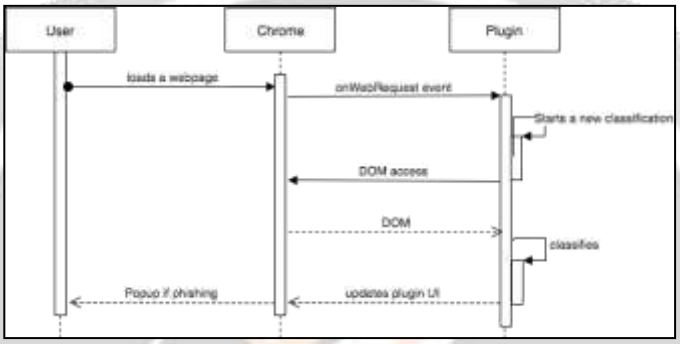

Fig-2: Working Sequence Diagram

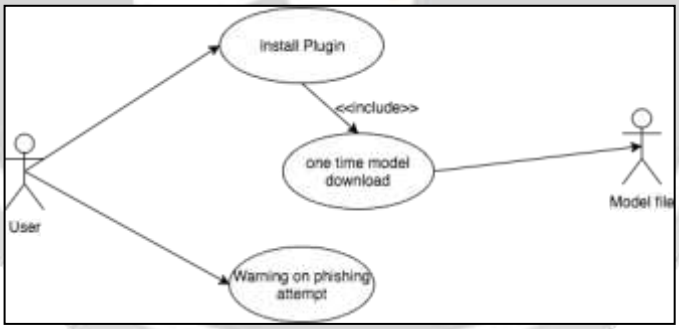

Fig-3: User Interface Diagram

There must be a simple and easy to use user interface where the user should be able to rapidly identify the phishing website. The input should be automatically taken from the web page in the current tab and the output should be clearly recognizable. Further the user should be interrupted on the effect of phishing.

As shown in Fig -4, Phishing detection browser extension detects the threats in the website and calculate safety percentage on 17 features and also give Phishing prediction about website.

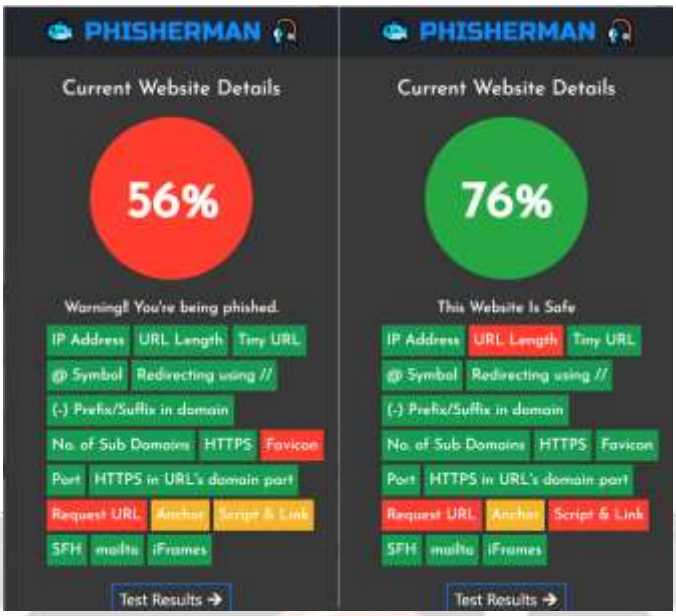

Fig -4: Actual Working Screenshots

5. CONCLUSIONS

Phishing is a manner to victimize via false e-mails and websites to steal someone's private information. Phishing prevents several from carrying their actions via the internet. Phishing website finding is important for the internet community since it has large impression on online transactions carried. Random Forest is an smart machine learning technique that freshly payed attention from researchers due to its speed and high categorization accuracy. The phishing website trouble has been investigated in this survey in which we developed a machine learning model to find out correlations between the characteristic and proceeds them from simple and effective pattern. In this study, we adopted classifier model that is used for detecting phishing websites in an intelligent, smart and automated way in real-time by using publicly available data-set. The execution of the proposed Random Forest classifier is preferably high in terms of classification quality, Fmeasure and AUC. Furthermore, our results showed that Random Forest is faster, robust and more faithful than the other classifiers. Random forest's runtime is quite fast & accurate, and it is able to detect phishing website in relation with the other classifiers.

6. REFERENCES

[1]. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine learning," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

[2]. Xiang, Guang, and Jason I. Hong. ―A hybrid phish detection approach by identity discovery and keywords retrieval.‖ In Proceedings of the 18th international conference on World Wide Web, pp. ACM, 2009

[3]. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no, pp. 949–952.

[4]. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1–6.