International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217354

Phishing Website Detection Using ML

Nikhil K*, Dr. Rajesh D S, Dhanush Raghavan
Department of Computer Science, Srinivas Institute of Technology, Mangalore, Karnataka, India

Article Info
Volume 7, Issue 4
Page Number: 194-198

Publication Issue :
July-August-2021

Article History
Accepted : 10 July 2021
Published : 15 July 2021

ABSTRACT

Phishing is one kind of cyber-attack , it is a most dangerous and common attack to retrieve personal information, account details, credit card credentials, organizational details or password of a client to conduct transactions. Phishing websites seem to like the relevant ones and it is difficult to differentiate among those websites. It is one of the most threatening that every individuals and organization faced. URLs are known as web sites are by which users locate information on the internet. The review creates warning of phishing attacks, detection of phishing attacks and motivate the practice of phishing prevention among the readers. With the huge number of phishing emails or messages received now days, companies or individuals are not able to find all of them.

Keywords : Phishing Detection, Decision Tree, Machine-learning

I. INTRODUCTION

Due to rapidly growing technology internet is now an integral part of our daily life. Most of activities in our daily life are depended on the use of the internet. Social networking sites have widely increased over the last few years. Due to the regular use of the internet, the users are under frequent and harmful threats; one of them is 'Phishing'. Phishing can be defined as impersonation of a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most wide spread cybercrime used today. There are countless domains where phishing attack may happen in online payment sector, webmail, and finances, file hosting or cloud storage networks and many others. The webmail and online payment sector has been attacked by phishing more than in any other industry. Phishing can be done through email phishing scams and spear phishing, which a user should be aware of the consequences and should not give their all-hearted trust on common security application. Machine Learning is one of the most efficient techniques to detect phishing as it removes drawback of various existing approaches.

The objectives, which is the most vital thing in proposed project, is to verity the validity of the website by capturing blacklisted URLs. To notify the user on blacklisted website through pop-up while they are trying to access the URL and a platform for an individual too check and validate the integrity of any URL they want to access.

II. LITERATURE SURVEY

In emerging technology, industry, which deeply influence today's security problems, has given a headache to many employers and home users.

Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. In these new times there are many security systems being enabled to ensure security is given the outmost priority and prevention to be taken from being hacked by those who are involved in cyber-offenses and essential prevention is taken as high importance in organization to ensure network security is not being compromised. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, etc. Customer face numerous security threats like cybercrime. Many cybercrime is being casually executed for example spam, fraud, identity theft cyber terrorisms and phishing. Among this phishing is known as the most common cybercrime today. Phishing has become one amongst the top three most current methods of law breaking in line with recent reports, and both frequency of events and user weakness has increased in recent years, more combination of all these methods result in greater danger of economic damage.

Phishing is a social engineering attack that targets and exploiting the weakness found in the system at the user's end. This paper proposes the Agile Unified Process (AUP) to detect duplicate websites that can potentially collect sensitive information about the user. The system checks the blacklisted sites in dataset and learns the patterns followed by the phishing websites and applies it to further given inputs. The system sends a pop-up and an e-mail notification to the user, if the user clicks on a phishing link and redirects to the site if it is a safe website. This system does not support real time detection of phishing sites; user has to supply the website link to the system developed with Microsoft Visual Studio 2010 Ultimate and MySQL stocks up data and to implement database in this system.

Phishing costs Internet user's lots of money. It refers to misusing weakness on the user side, which is vulnerable to such attacks. The basic ideology of the proposed solution is use to all the three-hybrid solution blacklist and whitelist, heuristics and visual similarity. The proposed system carries out a set of procedures before giving out the results. First, it tracks all "http" traffic of client system by creating a browser extension. Then compare domain of each URL with the white list of trusted domains and the blacklist of illegitimate domains. Further various characters in the URL is considered like number of '@', number of '-'and many more. Next approach is to extract and compare CSS of doubtful URL and compare it with the CSS of each of the legitimate domains in queue. This method will look into visual based features of the phished websites and machine-learning classifiers such as decision tree, logistic regression, random forest are applied to the collected data, and a score is generated. The match score and similarity score is evaluated. If the score is greater than threshold then the URL marked as phishing and blocked. This approach provides a three level security block.

Phishing is a dangerous effort to steal private data from users like address, Aadhar number, PAN card details, credit or debit card details, bank account details, personal details etc. The various types of phishing attacks like spoofing, instant spam spoofing, Hosts file poisoning, malware-based phishing, Man-in-the middle, session hijacking, DNS based phishing, deceptive phishing, key loggers/loggers, Web Trojans, Data theft, Content-injection phishing, Search engine phishing, Email /Spam, Web based delivery, Link Manipulation, System reconfiguration, Phone phishing, etc. are discussed in the paper. The recent approaches to prevent the attacks like heuristics approach, blacklist approach, fuzzy rule-based approach, machine learning approach etc. are also discussed and finally filtering all detection techniques

based on accuracy and performance proposed a framework to detect and prevent phishing attacks. A combination of supervised and unsupervised machine learning techniques is used to detect malicious attacks.

III. IMPLEMENTATION

A. Feature extraction

The feature extraction process is done from the URLs and corresponding binary values are given indicating whether the website is a phishing website or not. Below are the features that

We can extract for detection of fraud URLs.

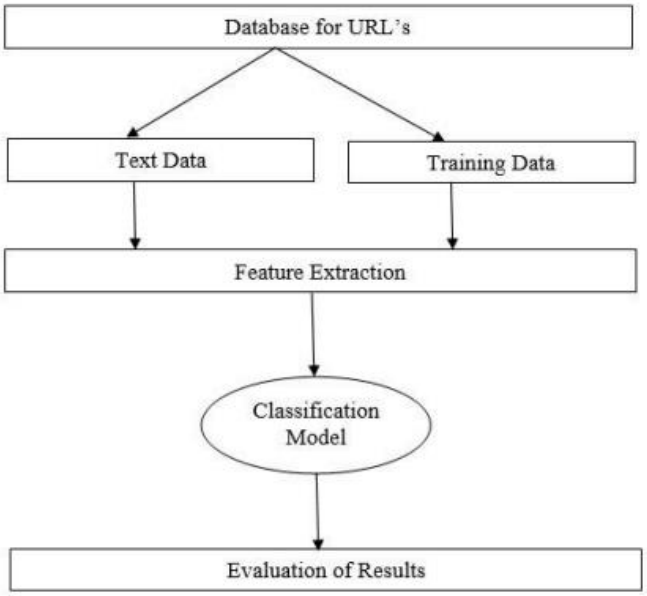
1. **IP address in URL:** If IP address present in URL then the feature is set to 1 else set to 0. Most of the legitimate sites do not use IP address in an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to collect sensitive information.
2. **'@' symbol in URL:** If @ symbol present in URL then the feature is set to 1 else set to 0.Hackers add special symbol @ in the URL leads the browser to ignore everything before the at the rate(@) symbol and the real address often follows the "@" symbol.
3. **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then the feature is set to 1 else to 0. This '-' symbol is rarely used in legitimate URLs. Phishers add hyphen symbol (-) to the domain name, so that users feel that they are dealing with a legitimate Webpage. For example, site is http://www.onlineamazon.com but phisher can develop another fake website like http://www.onlineamazon.com to trick the innocent users.
4. **Length of Host name:** Average length of the benign URLs is found to be a 25, If URL's length is more than 25 then the feature is set to 1 else to 0.
5. **HTTPS token in URL:** If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to fool users.
6. **URL redirection:** If "//" present in URL path then the feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be direct to another website.

B. Random forest Algorithm

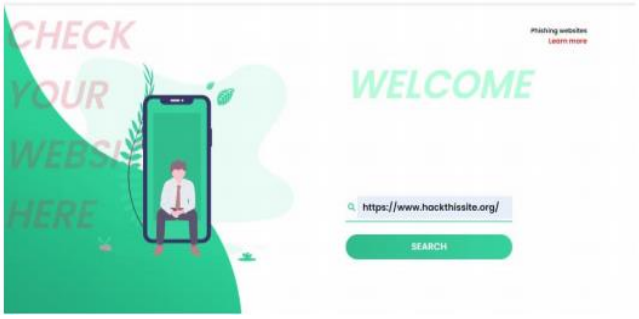
Random Forest is a machine-learning algorithm that belongs to the supervised learning technique; it can be applied for Classification and Regression problems. It is based on Phishing Website Detection using Machine Learning System Implementation the concept of Associative learning, which is a process of combining many different classifiers to solve a complex problem and to improve the efficiency of the model. As the name suggests, Random Forest is a classifier that contains a many number of decision trees on various subsets of the given dataset and takes the average to increase the predictive accuracy of that dataset. Instead of depending on one decision tree, the random forest takes the prediction from each tree and based on the maximum votes of predictions, it decides the final output.

C. Decision tree

Decision Tree is a supervised learning technique that can be used for classification and Regression problems, but mainly it is preferred for solving Classification problems. It is a tree structured classifier, which internal nodes represent the features of a dataset, branches indicate the decision rules and each leaf node indicates the outcome. In the Decision tree, there are two nodes, one is the Decision Node and other is Leaf Node. Decision nodes used to make many decision and have different branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.



The figure shows data flow diagram for the proposed system.



The above image is a home page we used in our system.

IV. RESULTS AND DISCUSSION

The proposed system enables the people to have a safe browsing and safe transactions. Its helps users to save their important credentials that should not be leaked. Our proposed system tells whether the given website is legitimate or not to users in the form of extension that makes the process of finding the truth about the website much easier. The results points to the efficiency with which our proposed system works to achieve the result using the hybrid solution of heuristic features, visual features and various approaches feeding these distinct features to machine learning algorithms. A particular challenge in this domains is that notorious hackers are constantly making new strategies to break into our defence measures. In order to get a well desired output we need algorithms that constantly learn and adapt to new examples and features of phishing URL's. And thus we use online learning algorithms. This new system can be designed to make use of maximum accuracy. Using different approaches altogether will improve the precision of the system, providing an efficient protection system. The drawback of this system is detecting of some minor false positive and false negative results. These disadvantages can be abolished by introducing much enhanced feature to feed to the machine learning algorithm that would result in much higher accuracy.

V. FUTURE WORK

Future work should focus on direct implementation of project to the chrome extension so that as the user clicks on the particular URL and if that URL is phishing site then the user gets a pop up warning message.

VI. REFERENCES

[1]. www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms

[2]. Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

[3]. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[4]. Rishikesh Mahajan (2018) "Phishing Website Detection using Machine Learning Algorithms"

[5]. Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[6]. S. Abu-Nimeh and T. M. Chen. Proliferation and detection of blog spam. Security & Privacy, IEEE, 8(5):42–47, 2010.

[7]. Jalil Nourmohammadi Khiarak (2017) "What is Machine Learning"

[8]. Tenzin Dakpa, Peter Augustine (2017) "Study of Phishing Attacks and Preventions"

[9]. Sadia Afroz, Rachel Greenstadt (2017) "PhishZoo: Detecting Phishing Websites By Looking at Them"

[10]. Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an Anti-Phishing Framework", 2019 5th International Conference On Advanced Computing & Communication System(ICACCS), pp. 1-6

Cite this article as :

Nikhil K, Dr. Rajesh D S, Dhanush Raghavan, "Phishing Website Detection Using ML", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 4, pp. 194-198, July-August 2021. Available at doi : https://doi.org/10.32628/CSEIT217354
Journal URL : https://ijsrcseit.com/CSEIT217354