International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Phishing Attack Detection and Prevention using Linkguard Algorithm

Harshali Dhanawde[1], Purva Dhainje[2], Swapnil Waghmare[3]

[1,2]Student,Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani, Maharashtra, India
[3]Assistant Professor, Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani, Maharashtra, India

---

Abstract - *The web technology has now become a wide range of platform where they make use of systems like mobile services and Internet of things. The data is clustered to the cloud-based platforms, and to access and operate this cluster of data we use various web applications, the web interface needs to be secure from technical attacks and social factor. The most forcibly deprived vectors in social engineering are phishing attacks. The user's sensitive information is collected by the attackers to imitate authentic web pages. The attackers equivocate the existing phishing defense mechanism based on URL's or page contents. As a robust basis to detect phishing attacks it is demonstrated that visual layout similarity has been considered in recent researches. In this paper, we aim to use machine learning techniques to affirm automated URL based phishing detection. To decide whether the page is legitimate or not, URL based phishing technique is used in this application. We generalize our solution and assess popular machine learning classifiers on their accuracy and factors affecting their results.*

Key Words: URL detection, detection using VStudio, machine learning, anti-phishing.

1. INTRODUCTION

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Typically, a victim clicks on a URL that appears to have been similar to a known contact or organization. That URL contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents. Many users unwittingly click phishing domains every day and every hour. The attackers are targeting both the users and the companies. The main reason is the lack of awareness of users. But security defenders must take precautions to prevent users from confronting these harmful sites. Preventing these huge costs can start with making people conscious in addition to building strong security mechanisms which are able to detect and prevent phishing domains from reaching the user. The main reason is the lack of awareness of users. But security defenders must take precautions to prevent users from confronting these harmful sites. Preventing these huge costs can start with making people conscious in addition to building strong security mechanisms which are able to detect and prevent phishing domains from reaching the user. In proposed System, which is design to prevent such threats. Extracts features from URL such as Length of host URL, No. of slashes, No. of dots in URL etc.

In this paper, we are focusing on detection of phishing websites by URL detection. In the earlier times, methods such as k-nearest neighbor, list-based approach, fuzzy logic, Phishzoo and other mining and classification approaches were used for detection but as the intensity of the attack grew over the times more sophisticated algorithms and techniques were introduced to detect and prevent the attack.

2. LITERATURE SURVEY

Phishing have various characteristics as well as it has various detection techniques and method. [5] The approaches for Phishing detection can be classified as whitelist, blacklist, content, visual similarity and URL based [5].

2.1 Whitelist

Whitelist is used as filter which blocks the Phish web pages to gain sensitive information from user. [6] The phishing pages are not blocked by the whitelist pass but they are further filtered using support vector machine classifier designed and optimized to classify threats like these.[6]

2.2 Blacklist

In this approach the requested URL is compared with predefined URLs. The drawback of this approach is that the

blacklist usually cannot cover all the phished sites as there are newly created fraudulent websites. Due to wide use of blacklist of websites against phishing, it is important to generate system which can generate updated black list of phishing websites.

2.3 Content based

[7] This technique generally downloads the content hosted at the URL and extracts the features from the content to identify Phish. These techniques require powerful website scraping to make sure the content is sufficiently retrieved. This detection can combine techniques that draws features from text of main index page and measure of visual similarity among websites to identify phishing attacks [7].

2.4. Visual similarity based

When two web pages are too similar then warning is raised. To determine similarities features like text pieces, images embedded in the page and overall visual appearance of the page are considered. Calculates the similarity between the target and legitimate page by comparing these features and computing a single similarity score.

2.5 based

[5]URL based detection can be done by using algorithm based, feature engineering based. The commonly used algorithms are Support vector machine, Naive Bayes, Decision tree, Random forest, Neural network. Commonly used features in URL based detection are IP address, URL length, No. of dots, special characters etc. [5].

3. EXISTING SYSTEM

Different studies and research related to phishing websites and their approaches were multi-tier that had been proposed which included classification for phishing URL filtering. These studies were displayed in different conferences of different journals. Weightage of message contents is based on an innovative method of extracting phishing URL features. [1]A multiple classification algorithm is used which includes SVM, AdaBoost, and Naive Bayes. These algorithms are divided into three tiers using 21 fixed yet different features [1]. Another classification algorithm that has two step procedure takes place here. It comes with a problem that addresses time consumption issues, complexity issues with a few overhead involved and performance issues which proves this method to be minimal or less at optimization[2] One of the primary methods used in this context is to recognize the attacks and then make use of feature extraction and then classification. The main limitation of this proposal is that there are too many features that are evaluated without considering whether they really are essential to identify phishing. Therefore, it could lead to unnecessary computational cost [2]. Blacklist-based and heuristic-based are the two approaches that fall under the detection techniques according to Institute of Research Engineers and Doctors, USA.[3] The blacklist-based approach maintains a database list of addresses (URLs) of those sites that are classified as malicious. If a user requests a site that is included in this list, the connection is blocked. The blacklist-based approach has the advantages of easy implementation and a low false- positive rate [3] One of the flaws is that it cannot detect phishing sites including temporary sites which are not listed in the database. The Multi-Label Classifier based Associative Classification (MCAC) Data mining approach is one of the methods is used for detecting phishing websites. According to International Journal of Advanced research and innovative ideas in education (IJARIIE) journal paper. The mediocre accuracy of these phishing websites is detected by the associative classification algorithm. [4] MCAC consists of three main steps which are Rule discovery, classifier building and class assignment. In the first step of this algorithm, rules are found and extracted by iterating over the training data set (historical websites features or data collected from various sources) .In this step, merging of any of the resulting rules that have the same antecedent (left hand side) takes place and are linked with different classes to produce the multi-label rules. Along with this, redundant rules are eliminated. The outcome of the second step is the classifier which contains single and multilabel rules. The last step involves testing the classifier on testing data set to measure its performance. In the prediction process, the rule in the classifier group which matches the test data features often fired to guess its type (class). The MCAC algorithm generate rules further that rules are sorted by using sorting algorithm. [4]. The difficulty in determining minimum confidence and minimum support when there is a large amount of data was one of the main problems that MCAC faced at that moment. more sophisticated algorithms to replace this which were more accurate and had lesser time complexity were the other issues that MCAC had to face.

4. IMPLEMENTED TECHNIQUE

Phishing attacks through URLs is one of the major issues faced by the Internet Community because of the online transactions performed on a day-to-day basis. However, there are anti-phishing tools available which can help users to detect phishing attacks and prevent them. Malicious URLs

are detected by heuristic based Phishing detection System and specify the reason for classifying a URL as phishing which will help the user to aware of such malicious as well as suspicious URLs. When the user enters a URL ,the system will provide heuristic for user to select and apply on input URL is malicious or legitimate .It stores all the input URLs in the database which can be retrieved for future use .Testing is done on all legitimate websites as well as malicious websites which are collected from Phish tank .The testing is done on combinations of multiple heuristics also on individual heuristics to ensure the efficient functionality of system. From set of URLs being tested, majority of the URLs have been classified as correctly by the system. Evaluation of the system is done using a confusion matrix which lists the true positives, true negatives, false positives and false negatives Once all this information is collected ,the precision and the recall is calculated for the system .The precisions ad recall varies accordingly based upon the heuristics selected by the user .For better precision and recall ,the false negatives and false positives can be reduced which will improve classification accuracy. The discriminative classifier algorithms such as SVM, Decision tree, Boosting can be used to predict the URL category by training huge amount of the data extracted from the datasets. Here we are using Decision tree algorithm for classification.

5. SYSTEM ARCHITECHTURE

Implemented system detect phishing sites efficiently. A lot of work and researches have been done related to above topic the implemented system focuses on detecting phished URLs by extracting features with the help of factors like page rank, no of slashes and dots used in URL. WEKA is used to determine the accuracy and performance of various classification algorithms. A lot of journals and research papers have been read to decide the classification algorithm. The idea that we are putting here is to detect phishing URLs and improve the efficiency by using decision tree algorithm with help of visual studio for better analysis.
Table 1 shows the list of features that we have considered for feature extraction.

Table -1: List of Features used for feature extraction

Sr No	Features
1	IP Contains
2	Length of URL
3	Suspicious char
4	Prefix Suffix
5	Dots
6	Subdomain
7	Slash
8	http

Decision tree algorithm is implemented using Visual Studio. A Decision Tree can be considered as an improved nested-if-else structure. Each feature will be checked one by one. Decision Tree uses an information gain measure which indicates how well a given feature separates the training examples according to their target classification. The name of the method is Information Gain**.** High Gain score means that the feature has a high distinguishing ability. Because of this, the feature which has maximum gain score is selected as the root. In the training phase, dataset is divided into two parts by comparing the feature values. [8] The database named "Phishtank" is used to gather information, since there is huge amount of data is available for processing [8].

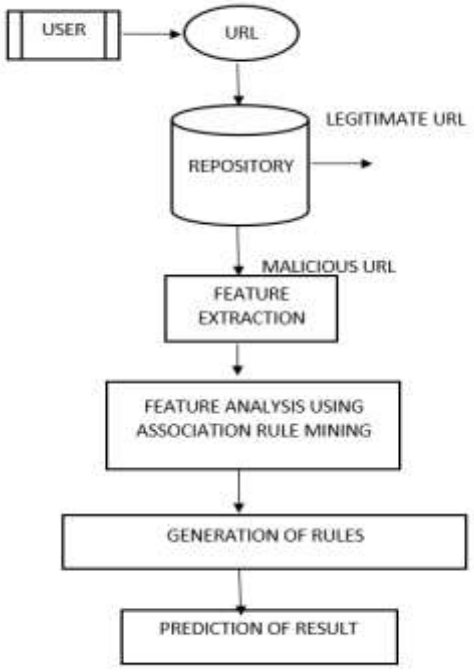

Fig -1: System Architecture

6. RESULT AND ANALYSIS

Implemented system has comparatively more accuracy than other classification algorithms like Naïve Bayes, SVM, Random forest etc. This system not only detects the requested URL is Legitimate or Malicious by checking it from the database which contains lots of URLs classified into two subsets viz Legitimate and malicious but also examines the new URL which is not there in database. After identifying whether it is Phished or Legitimate system will store new URL in database for further use. When user enters query, it will check whether it is legitimate or not. If it is Legitimate then it will direct user to the website otherwise show the

popup indicating its Phished URL. Also it can test the list of URLs and visualize it in the graphical form.

Table-2: Evaluation between Existing and Implemented Technique

Features	Existing Technique	Implemented Technique
URL based Detection using Link guard Algorithm	No	Yes
Detection of Phished URL excluding from the data in the Database	No	Yes
Using CSS layout in order to lessen the Accuracy	Yes	No

Fig -2: Query box for user to enter URL

Fig-3: Legitimate URL

Fig -4: Query box for user to enter URL

Fig -5: Phished URL

7. CONCLUSION

In this paper we've discussed about implemented system, using decision tree method as it gave an optimum result compared to other algorithms namely Naïve Bayesian, SVM and Random Forest. The accuracy was calculated which proved that this method is a good metric for unbalanced data. The dataset needed for the entire procedure was gathered from Phish tank and since there was a large amount of data to process, parsing was performed primarily. Parsing is done to analyze feature set. The dataset was divided into 0's and 1's where 0 indicates phished website and 1 indicates legitimate websites.

This study is useful for the users to prevent themselves from phishing activities which lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details and passwords.

8. REFERENCES

[1] International Journal of Advanced Computer Technology (IJACT), "A Review of Various Techniques for Detection and Prevention for Phishing Attack".

[2] IEEE 2017 - Feature selection for machine learning based detection of phishing websites "http://ieeexplore.ieee.org/abstract/document/8090317/?reload=true"

[3] Heuristic-based Approach for Phishing Site Detection Using URL Features- Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015 Copyright © "Institute of Research Engineers and Doctors, USA. All rights reserved."

[4] Prof.T.BhaskarAher Sonali, Bawake Nikita , Gosavi Akshada Gunjal Swati ' Detection of Website Phishing Using MCAC Technique Implementation', "http://ijariie.com/AdminUploadPdf/Detection_of_Website_Phishing _Using_MCAC_Technique_Implementation_ijariie1807.pdf"

[5] DEIM Forum 2019 G2-3 - A Survey of URL-based Phishing Detection "https://db-event.jpn.org/deim2019/post/papers/201.pdf"

[6] 2012 Seventh International Conference on Availability, Reliability and Security - A personalized whitelist approach for phishing webpage detection "https://sci-hub.tw/https://ieeexplore.ieee.org/document/6329190''

[7] High Performace content based phishing attack detection "https://www.uab.edu/cas/thecenter/images/Documents/High-Performance-Content-Based-Phishing-Attack-Detection.pdf"'

[8] Phishtank - https://www.phishtank.com/