IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 7, PP 55-57

Detection of Phishing Site Using Efficient Approach

Dhanashree Pawar[1], Dhanalakshmi Sherbhai[2], Akshata Shelar[3], Neha Singh[4]
[1](Student, Computer Engineering, Atharva College of Engineering, Mumbai University, India)
[2](Student, Computer Engineering, Atharva College of Engineering, Mumbai University, India)
[3](Student, Computer Engineering, Atharva College of Engineering, Mumbai University, India)
[4](Assistant Professor, Department of Computer Engineering, Atharva College of Engineering, India)
Corresponding Author: Dhanashree Pawar

Abstract: *Phishing is an online criminal act that occurs when a malicious webpage impersonates as legitimate webpage so as to acquire sensitive information from the user. Phishing attack continues to pose a serious risk for web users and annoying threat within the field of electronic commerce. This paper focuses on discerning the significant features that discriminate between legitimate and phishing URLs. These features are then subjected to associative rule mining. The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. Analyzing the knowledge accessible on phishing URL and considering confidence as an indicator, the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL. The objective of the proposed system is to provide efficient approach for the detection of phishing site.*

Keywords – *Phishing attack, C4.5, Naïve Bayes*

I. Introduction

With the recent growth of the Internet environment and diversification of available web services, web attacks have increased in quantity and advanced in quality. Phishing is a type of social engineering attack that targets a user‘s sensitive information through a phony website that appears similar to a legitimate site.This project aims at detecting phishing sites made by the phishers who steal personal user data to conduct illicit activities. We will extract features from the submitted URL by the user. These features will then be given to decision tree algorithm to classify the site as phishing or legitimate. Also, ranking of the sites will be considered while classifying the site as phishing or legitimate.

II. Basic Concept

Phishing sites are the major attacks by which most of internet users are being fooled by the phisher. The replicas of the legitimate sites are created and users are directed to that website by luring some offers to it. There are certain standards which are given by W3C (World Wide Web Consortium), based on these standards we are choosing some features which can easily describe the difference between legit site and phish site.

We are proposing a model to determine the phishing sites to Safeguard the web users from phisher. The features of URL are taken into consideration for detecting valid or invalid phish. Also the comparison between two approaches i.e. Naïve Bayes Algorithm and C4.5 Algorithm is shown with the time and space complexity analysis.

III. Problem Statement

In modern times as the techniques for Phishing Detection have advanced, various methods present some advantages as well as issues. Data mining techniques have been used in Phishing Detection since ancient time and its usage can never go obsolete. Hence, there are many systems implemented in this field As a result, we need a system with,

1. Appropriate methodology
2. Less processing time
3. Good value of evaluation metrics

The proposed system focuses on yielding accurate results regarding the decision about Phishing site or legitimate site, improving the value of various evaluation metrics, less processing of time and fast retrieval of data. Heuristic based approach along with decision tree algorithm is used in the system to enhance the accuracy of the system. For Classification C4.5 algorithm is used. This algorithm is the incremented version of ID3

algorithm[4]. For heuristics various URL features such as Primary Domain, length of URL, suspicious characters, DNS record etc. are considered.

IV. Proposed System

The proposed system aims to increase the accuracy of phishing detection systems that aim to differentiate between phishing sites and legitimate sites with the help of URL features using Heuristic Approach [10]. The proposed technique extracts features in URLs of user-requested pages and applies those features to determine whether a requested site is phishing site. This technique can detect phishing sites that cannot be detected by blacklist-based techniques; therefore it can help reduce the damage caused by phishing attacks.
The System Flow diagram of the proposed system is depicted as follow in figure:

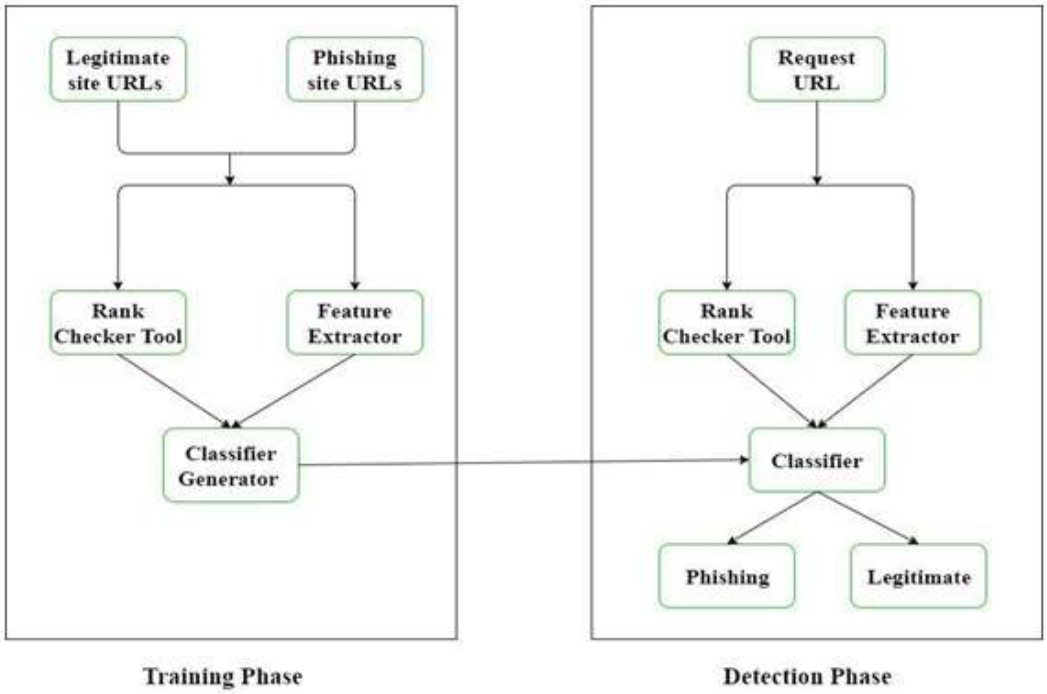

Fig. 4.1: System Flow Diagram

4.1 Method Overview

In this System, admin will upload a dataset which will contain a list of URLs (List of Legitimate and list of phishing sites). The admin will train the system using C4.5 and Naive Bayes Algorithm. Using C4.5 algorithm, the admin will generate rule set which will be saved onto the user system. After training the system, admin tests the system by giving input as URL to the system [5].The system will predict whether the site is phishing or legitimate site. The Result will be stored in the database retrieved. The extracted features are given as input to previously generated classifier. This classifier will predict whether the URL is phishing or legitimate site [10].

Using Naive Bayes Algorithm, the admin trains the system; user can test the system by giving input as URL to the system. It predicts whether the URL is phishing or legitimate site based on Naive Bayes Algorithm. Result will be stored in the database. The admin can compare the accuracy of the system by comparing both the algorithm results in graphical format.

V. Conclusion

The Previous System Used Listing Method And Naive Bayes Classifier To Classify Phishing Sites. In The Proposed System A URL Based Phishing Identification Technique Is Used That Employs URL Based Features. To Make Feedback For Each URL More Meaningful, A Heuristic System That Gives Potent URL Classification Feedback Is Implemented. The Method Combines URL Based Features Used In Previous Studies With New Features By Analyzing Phishing Site Urls. Additionally, Classifiers Are Generated Through Several Machine Learning Algorithms And It Is Determined That The Best Classifier Is C4.5. Various Heuristics Are Used To Obtain A Classifier That Would Be Able To Achieve High Accuracy, While Maintaining A Minimal False Positive Rate. Training The Machine Learning Algorithm Enables The Classifier To Learn New Trends In The Characteristics Of Urls Over Time In A Quick Manner. The Proposed Technique Can Provide Security For

Personal Information And Reduce Damage Caused By Phishing Attacks Because It Can Detect New And Temporary Phishing Sites That Evade Existing Phishing Detection Techniques, Such As The Black-List Based Technique And White-List Based Technique Along With Provision Of Optimum Time And Space Complexity.

References

[1] Altyeb Altaher, "Phishing Website Classification using Hybrid SVM and KNN Approach", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017

[2] Rahul Patil, "Hybrid model to detect Phishing sites using Clustering and Bayesian Approach", International Conference for Convergence of Technology – 2014

[3] A. P. Deore, "Phishing Detection Information Identification using Support Vector Machine" , Internation Journal of Innovation in Engineering and Technology, 2015

[4] Badr HSSINA, "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications

[5] Zhu Xiaoliang , Wang Jian, "Research and application of algorithm C4.5 on decision tree", 2009 International Conference on Test and Measurement

[6] Sophie Gastellier-Prevost, Gustavo Gonzalez Granadillo,"Decisive Heuristics to differentiate Legitimate from Phishing sites", Network and Information Systems Security (SAR-SSI) Conference, 2011

[7] N. Badadh, S. More, N. Puri, "An Efficient Approach To Detecting Phishing Web Using K-Means And Naïve-Bayes Algorithms With Results", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014

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

[9] RakeshR, Kannan A, Muthurajkumar S, "Enhancing the Precision of Phishing Classification Accuracy using Reduced Feature Set and Boosting Algorithm", 2014 Sixth International Conference on Advanced Computing(ICoAC)

[10] Jin-Lee Lee, Dong-Hyun Kim, "Heuristic-based Approach for Phishing SiteDetection Using URL Features", Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015
Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishingdetection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

[11] V. Ramanathan, H. Wechsler, "Phishing website detection using Latent Dirichlet Allocation and AdaBoost," IEEE International Conference on Intelligence andSecurity Informatics. Cyberspace, Border, and Immigration Securities, Piscataway, NJ, pp. 102–107, 2012.

[12] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator, " in Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, ser. AlCCSA '08. Washington, DC, USA: IEEE Computer Society, pp. 840-843, 2008.