URL Based Analysis for Phishing Detection Using Data Mining

Nikita Gawade
Information Technology
Shah and Anchor Kutchhi
Engineering College
Mumbai,India

Sayali Mundekar
Information Technology
Shah and Anchor Kutchhi
Engineering College
Mumbai, India

Nilam Vare
Information Technology
Shah and Anchor Kutchhi
Engineering College
Mumbai,India

Ruchi Gada
Information Technology
Shah and Anchor Kutchhi
Engineering College
Mumbai,India

Prof.Smita Bansod
Information Technology
Shah and Anchor Kutchhi
Engineering College
Mumbai,India

***Abstract*— Phishing sites are those sites that are designed by deceptive people to look same as legitimate sites with one or more different features which cannot be noticed by naive users. These phishing URLs seems like an original website of any organization such as bank, institutes, etc. In phishing, attacker directly attacks on user's personal information. Phishing is a cybercrime in which victims are targeted by email, telephone or text message by stating that it is a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. Data mining technique can be used for classification of different data which was used for detection of phishing websites as compared to non-technical approaches like blacklist-based detection techniques. Different classification algorithms can be used to predict different phishing URLs. In this paper we analyse url features using random forest algorithm to detect phishing websites. The study analyses over legitimate and phishing URLs collected from phishtank. It includes classification of different URLs using random forest classifiers .**

Keywords— Phishing, URL, Website, Data mining, legitimate,Random Forest.

I. INTRODUCTION

Most internet users have encountered phishing in the form of emails pretending to come from bank or other business but originating from a malicious source and design to persuade the recipient to handover personal information such as credit card details. Phishing is done by gaining the trust of email receiver such that they will convinced that the message is coming from a bank or a company about the facilities or services they want, for example, a mail from a company or a bank that to go to a link or download afile related to the services or for any other use . There are variety of types that falls under phishing to get important information such as handover sensitive information in which attacker trick the user into revealing important data often a username and password that the attacker can use to breach a system or account and download malware in which there will be a lot of junk files due to which their computer will get affected by the virus. Spear phishing targets a specific person or enterprise, as opposed to random application users. It's a more in-depth version of phishing that requires special knowledge about an organization, including its power structure. Spear phishing is a attack where people get manipulated psychologically in which a perpetrator, disguised as a trusted individual, tricks a target into clicking a link in a spoofed email, text message or instant message. As a result, the target unwittingly reveals sensitive information, installs malicious programs (malware) on their network or executes it.

In response to this increase in phishing attacks, phishing detection techniques need to be applied. Detection of phishing can be done by using data mining algorithms and by using machine learning. Typical phishing detection techniques include the blacklist-based detection method and heuristic-based technique. The blacklist-based technique maintains a uniform resource locator (URL) list of sites that are classified as phishing sites and heuristic based detection technique analyses and extracts phishing site features and detects phishing sites using that information.

This study aims towards enhancing software classification algorithm, which will compete against phishing attacks. The solutions which were given previously depends on analyzing the content of websites such as PHP content, HTML content etc but this study mainly focuses on URLs of the websites.

II. LITERATURE SURVEY

Now a days Phishing attacks has increased so much that it has become serious problem for every individual. This attack uses victims credentials for different malicious activities. Many anti-phishing techniques have been proposed To overcome these malicious activities. By the use of blacklist approach which are used by various web browsers, many antiphishing tools has been launched such as Net craft AntiPhishing Toolbar, eBay Toolbar, SpoofGuard, TrustWatch etc. Out of these tools, tools like SpoofGuard and TrustWatchfalsely indicates some of the original sites as phishing sites. Data mining algorithms have been used by many researchers for phishing detection. Data mining techniques can detect phishing websites in real time as Data Mining algorithms have been proved to be useful..

[2] In one of the paper lexical URL have been used to identify phishing and legitimate website. URLs were lexically analyzed to extract characters and features of URL which called as tokens that appear more frequently in phishing and legitimate URLs and then these tokens are used in the binary format (e.g if token found then 1 otherwise 0). In this way the token dictionary for both phishing and legitimate URLs is constructed. A statistical classifier is used that takes

Unclassified URLs as a input and predict the output in binary class (e.g Phish or Legit). This is done by the classifier by parsing the URL and dividing it into tokens and find the number of presence of each token also with its relevant class. This study empirically confirms that URLs contain more information about the phishing sites and this token based analysis resulted in very good classification accuracy of 97%.In this way classification data mining algorithms can be used to detect phishing websites.

[1]In other paper it uses C4.5 data mining algorithm which is a ideal data mining technique which can correctly identify phishing websites. This analysis has been used 750 URLs as a training dataset and 300 URLs as a testing dataset. Training dataset is used to train the algorithm J48 which is an implementation of C4.5 data mining algorithm. After training the classifier has been generated this used to make prediction. True positive rate, True negative rate, False positive rate, False negative rate, Success rate, Error rate and Accuracy are calculated after testing process. Result shows C4.5 has an accuracy of 82.6%.

[3]In other paper, Five attribute selector algorithm have been used (PCA, CR, CFS, IG and GR) to analyze the feature set. It has been observe that accuracy rate of CFS is 95% ,IG is 96% and PCA is 95%.Then it has been compare classification algorithm including Bayesian Network (BN), J48 Decision Tree, Random Forest (RF) and Random Tree(RT). After that it has been conclude that feature subsets obtain by the information gain algorithm have higher stability, RF algorithm has better accuracy, but its take more time to construct the model, However, RT's performance can have better accuracy and time complexity. Based on, this analysis it is identified that RT is best classifier for phishing detection.

[4]One of the research uses machine learning algorithms such as support vector machine (SVM), naive Bayes, decision tree, k-nearest neighbor (KNN), random tree, and random forest. For comparison they gather 3,000 phishing site URLs from Phishtank and 3,000 legitimate site URLs from DMOZ. As a result of the experiment, it has been determined that the best machine learning algorithm, random forest, used URL features. It showed a high accuracy of 98.23% and a low false-positive rate. This classifier detected more than 98.23% of phishing sites.

III. PROPOSED METHOD

Datasets of both phishing as well as legitimate websites has been collected from Phishtank. Further classification of websites into phishing and legitimate websites is performed using Random Forest Classification algorithm. Classification of phishing websites is done which is followed by URL based analysis of websites.

A URL is a reference to a web resource that specifies its location on a computer network which is used to retrieve it.We define features of each component of the URL which will be used for phishing site detection.

The URL is composed of the protocol identifier,
Sub domain, primary domain, top-level domain (TLD), and path domain or requested content.. In this study of URL structure for phishing detection, the subdomain, primary domain, and TLD are collectively referred to as the domain.

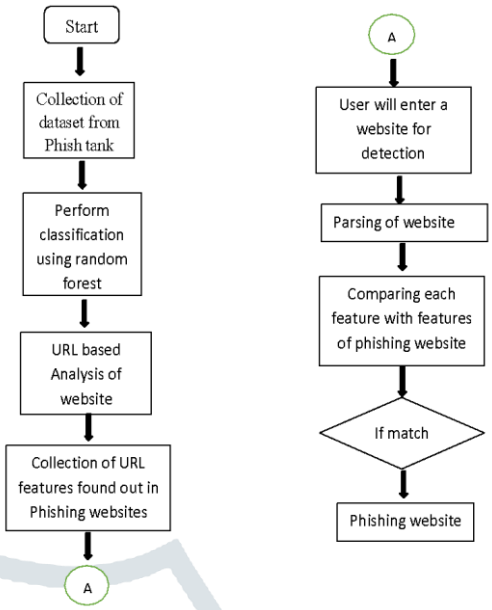

Fig 1-Workflow Diagram

URL based Analysis:

Fig. 2 depicts the individual components of the URL. The protocol identifier indicates the name of the protocol to be used to fetch the resource eg. HTTP, HTTPS, FTP, Grofer. Protocols are of various types and are used in accordance with the desired communication method.

Fig 2 . URL Structure with components

The subdomain is a secondary domain given to the domain and has various types depending on the services provided by the domain page. In the aforementioned example www is the subdomain. The domain name here in our example URL nyszone is a unique identification string that defines a realm of administrative authority or control within the internet. The domain is the most important part of a URL. The first-level set of domain names are the top-level domains (TLDS) such as the prominent domains com, info, net, edu and org and the country code top-level domains (ccTLDs), The second and third-level domain names are typically open for reservation for end users who wish to run their own websites.

RANDOM FOREST Algorithm :

Random Forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. All trees in the forest have the same distribution. This algorithm can handle a large number of variables in the dataset; however, it lacks reproducibility because the process of forest building is random.

In this phase we are going to develop an application which will be using URL Features as Identified from the phishing websites in Phase 1. This app will have a Database of these phishing URL features. When a user enters any URL for detection the app will parse the website. The app will now compare every parsed feature with phishing features and if identified any it will be classified respectively as phishing or legitimate.

IV. CONCLUSION

Phishing is the fraudulent attempt to obtain sensitive information such as user-names, passwords and credit card details by disguising as a trustworthy entity in an electronic communication. Typically carried out by email spoofing or instant messaging, it often directs users to enter personal information at a fake website, the look and feel of which are identical to the legitimate site.

Phishing sites cannot be easily identified by naïve users so many Anti-phishing techniques have been evolved. Cyber criminals indulge in the practice of phishing. For evaluating the phishing websites, we can use technical and non-technical approaches. In technical approach, Blacklist and heuristic method is used to detect phishing sites and innon-technical approach, we use legal solution and training people.

Phishing is a negative use of social engineering techniques to deceive users. Users are often lured by communications persuading to be from trusted parties such as social web sites, auction sites, banks, online payment processors or IT administrators

The annual worldwide impact of phishing could be as high as US$5 billion. Many researchers have worked on phishing detection techniques using data mining algorithms such as C4.5, J48, blacklist, heuristics, etc. Various techniques have been implemented to control the phishing attacks. Thus an efficient phishing detection mechanism needs to be designed to provide security to naïve users from data theft and credentials accession by hackers which might lead them to a great financial loss. The phishing detection mechanism should also prevent unauthorized access to sensitive information.

REFERENCES

1]. A. Priya, Er. Meenakshi, "Detection of Phishing Website Using C4.5 Data Mining Algorithm", 2017 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology RTEICT, May 19-20,2017, India.

[2]. Mahmoud Khonji, Youssef Iraqi, Andrew Jones," Lexical URL Analysis for Discriminating Phishing and Legitimate Websites", C[EAS' 11 September 1-2,2011, Perth, Western Australia, Australia copyright 2011 ACM 978-1-4503-0788-8/11/09.

[3]. Zhao Zhang, Qinggang He, Bailing Wang, "A Novel Multi-Layer Heuristic Model for Anti-phishing", ICIE '17, August 17-18,2017, Dalian Liaoning, china 2017 Association for Computing Machinery, ACM ISBN 978-1-450-5210-9/17/08.

[4].Maher Aburrous, M. A. Hossain, KeshavDahal,Fadi Thabtah, "Predicting Phishing Websites using Classification Mining Techniques withExperimental Case Studies", 2010 Seventh International Conference on Information Technology.

[5].Anindita Khade,Dr. Subhash K Shinde,"Detection of Phishing Websites Using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 12, December – 2013 ISSN: 2278-0181

[6]. Samruddhi Yadav,Smita Bansod" Comparative analysis of anti-phishing method",2015-2016.

[7].S. Garera, N. Provos, M. Chew, and A. D. Rubin. "Aframework for detection and measurement of phishingattacks." In Proceedings of the 2007 ACM workshop onRecurring malcode, WORM '07, pages 1–8, New York, NY, USA, 2007. ACM.

[8].Chen Y S, Yu Y H, Liu H S, et al.2014. Detect phishing by checking contentconsistency[C]//Information Reuse and Integration (IRI), 2014 IEEE 15thInternational Conference on. IEEE, 2014: 109-119.