Improved Method Text Feature Extraction for Detection of Phishing Website

Ganesh S. Gangthade, Shyam R. Gote, Sachin P. Khengte, Mukund K. Kharade, Prof. Rashmi Tundalwar

ABSTRACT
Phishing is a technique of gaining personal information of users from various websites. Sometimes it redirects the user to phish webpage to gain information of user like username, password, account and credit card details etc. Our main ambition here is to design system to provide safeguard to users against phishing attacks. Our work is mainly focuses on use of terms and URLs from web page to detect possible phishing patterns from web pages of phishing websites. Process initiates with parsing of web page to extract plain text terms and URLs. Further detected terms are fed to TF-IDF and URL weighting system to identify importance of each detected term. Later search engine lookup is carried out for most important terms which help to detect possible victim URLs for given input website. Finally WHOIS lookup is used to compare registration details of websites to correctly categorize website as phishing or legitimate one.

Keywords: Phishing detection, K-means ,Naïve Bayes Algorithm, URL features

I. INTRODUCTION
Phishing is a relatively new internet crime in comparison with other forums, e.g., virus and hacking. Due to the requirement of internet users to facilitate them for 24/7 for banking, housekeeping activities and various many more needs, phishing attacks keep growing. More and more phishing web pages have been found in recent years in an accelerative way (Fu, et al., 2006). Its impact is the breach of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds.

A phishing website is a broadly launched social engineering attack that attempts to defraud people of their personal information including credit card number, bank account information, social security number, and their personal credentials in order to use these details fraudulently against them (James, 2006). Phishing websites use a number of different techniques to hide the fact that they are not authentic including overwriting or disguising the true URL shown in the browser, overlaying the genuine web site with a crafted pop-up window, drawing fake padlock images on top of the browser window to give the impression that SSL is enabled, and registering SSL certificates for domain names similar to the real organization etc.

Different types of techniques are used for phishing detection and avoidance websites attacks. It mainly includes users browser based dynamic security, predefined rules for web page creation by Website Company, visual and DOM tree similarity based approach and comparing URLs with blacklisted sites. Even though so many approaches are proposed there is need of approach which not only detects phishing websites but also let rethink web page phisher before using contents of legitimate websites.

II. LITERATURE SURVEY

To detect and prevent various kinds of phishing attacks, there are many different preventive strategies and
III. PROPOSED SYSTEM

In this project, a fast and accurate approach is proposed to detect phishing web. Our approach determines whether a webpage is a phishing web or a legitimate one, based on its URL and webpage features, and is merely a combination of K-Means and NB. The K-Means classifier used to detect the URL is that K-Means is a rapid detection method for classification and URL features can be easily acquired. If the K-Means classifier cannot judge the given web’s legality definitely, the Naive Bayes classifier is used to detect it based on its webpage features. Also our approach may work together with a blacklist-based method to provide a better protection.

URL Feature Set: The main working of the model depends upon what features are to be used in the dataset to detect the phishing attack. After studying W3C standards we have chosen following four features which can effectively determine the phishing attacks:

1. IP Address in URL: The domain needs to be registered in order to obtain a specific URL address for the web site but the phishing sites do last only for few days hence the phisher may not register the web site. Legit sites have their domain registered and have the URL address. This will help us to determine the phishing site.
2. Dots in URL: The dot in the URL represents the existence of the sub-domain in the URL. Some phisher may use the sub-domain to look the site address as the legit site hence causing the user to mislead to phish site.
3. Suspicious Characters: The Phisher will use some special characters other than alpha-numeric character to trick the user. Special characters used may be ‘@’, ‘&’, ‘.’, and ‘.’ in the web URL to create the pattern of the legit URL which the user easily click on.
4. Slashes in URL: The slashes in the URL shows existence of sub-folders in it. The sub-folders are added to hide the information in the web-pages.

Algorithm:
Step 1: Given a web P, extract its URL identity and generate features.
Step 2: Classify P by K-Means classifier and return result (+1, -1 or 0).
Step 3: If result=+1 or -1, output the phishing label.
Step 4: If P has not a text input, output the phishing label (1).
Step 5: Extract its webpage identity and generate features.
Step 6: Classify P by Naive Bayes classifier and output the phishing label.

IV. CONCLUSION

In this paper, we evaluated two phishing detection system mechanisms out of which one is dependent on URL features of web-sites and Visual Features of web-sites. We have created a system which is a trail of combination of these two systems and using base techniques given by them. Application of clustering on this system generates the output faster but by compromising with the accuracy of results.

V. REFERENCES

[4] Ee Hung Chang, Kang Leng Chiew, San Nah Sze and Wei King Tiong, “Phishing Detection via Identification of
Website Identity’, Faculty of computer science and Information Technology University Malaysia Sarawak, IEEE 2013.

