Classification of Phishing Website Based on URL Features

S.Aarthi [1], Narsepalli Vamsi Kishan [2], V.Surya Teja [3], N.V.Harsha Vardhan Gupta [4]
[1] Assistant Professor, Dept. Of Computer Science Engineering, SRM Institute of Science and Technology, Chennai, India.
[2, 3, 4] Dept. Of Computer Science Engineering, SRM Institute of Science and Technology, Chennai, India.

Abstract – Phishing is a type of Internet fraud that seeks to acquire a user's credentials by deception. It includes stealing of passwords, Mastercard numbers, bank account details, other confidential information. Phishing messages sometimes take the shape of faux notifications from banks, providers, e-pay systems and alternative organizations. Already a few methods are proposed to counter phishing attacks, but none of them are perfect solutions. One of the best ways to prevent phishing is through data mining by accessing data of a particular website. In this paper, we propose an algorithm called URL Mining algorithm to analyze a URL in different ways. Based on different parameters used we extract URL features of the website and classify it as a phishing website or genuine website. This strategy provides security from phishing websites by 98.5% compared to other methods.

Index Terms – Phishing Website, Data Mining, URL Mining Algorithm, URL Features.

1. INTRODUCTION

Phishing is a criminal activity to steal consumers' personal identity data and financial account credentials. Fraudulent use spoofed e-mails imitating to be from legitimate businesses and agencies, designed to lead consumers to counterfeit Web sites that trick recipients into divulging financial data such as usernames and passwords. They try to steal credentials directly, often using systems to intercept consumers online account user names and passwords and to corrupt local navigational infrastructures to misdirect consumers to counterfeit Web sites. Just during the month of Oct-Dec, 2018, the no. Phishing attacks detected are 1,38,328.

In 2018, 78% of Indians and 83% of Americans regularly shopped online. With an increase in technology most of the financial and government organizations have extended their online services to their clients using smart phones, increasing number of people are depending on online services to shop, check their banking account, pay their bills. While such activities had an important impact on the world economy, such large dependencies on online financial services increases security risks for both customers and financial institutes. Without knowledge of the customers they are becoming vulnerable to the phishing attacks.

So as to help people detect these type of phishing sites which slip through the classification of search engines like Google, Mozilla Firefox etc. we propose this paper. In this we use URL Mining algorithm to detect the features of the URL. Based on the features detected we compare them with the standard URL features and classify whether it is a phishing website or not.

2. RELATED WORK

2.1 A detection model proposed by Mohammed et al. and calculated the detection error-rate d by the associative classification algorithms. The results of the research is that C4.5 has an average error-rate of 5.76%.

2.2 Aburrous et al. Proposed a system for phishing detection in e-banking. They proposed a model based on fuzzy logic combined with data mining algorithms to examine the techniques by describing the phishing website aspects and by categorizing the phishing types. By using 10-fold cross-validation, they achieved 86.38% classification accuracy, which is very low.

2.3 Arade et al. Proposed system to compare the addresses in the database of the proposed system and the webpage address. In this study, the problem is it may consider legitimate webpages as phishing webpages.

2.4 Shahriar & Zulkernine proposed a model for detecting phishing webpages using the reliability of suspected pages. In their study, a finite state machine is proposed to assess webpage behavior by tracing the webpage form the submission as well as from the corresponding responses.

2.5 Ajlouni et al. proposed the system with MCAR by observing the features from Aburrous et al. work and achieved 96% accuracy in classifying the webpages, but they did not give any explanation on how they were extracted by using the MCAR algorithm.

2.6 Ramesh proposed a system which distinguish all the direct and indirect links related to the webpage suspected. The indirect page links are taken out from the search engine result but the direct links are taken out from the page content itself. In order to get the result they also use Third party domain.

2.7 Zhang et al. used Sequential Minimal Optimization classifier with five features to distinguish Chinese phishing websites. The limitation of this approach is that the extracted

features are only for detection of phishing webpages with Chinese language. Li et. al. used transudative support vector machine to detect and classify phishing web pages. They extract the features of the web page image to reflect the characteristics of web pages absolutely.

3. PORPOSED MODELLING

The proposed system uses URL Mining algorithm to analyze website URL's. From the system architecture in Fig.1 we can see system is divided into three modules namely classifier, feature extraction and feature analyzer. It follows the following URL mining algorithm.

1. Access webpage in the browser.

2. If it is from an authentic website, not a phishing webpage(based on the addresses already saved).

3. If it is a suspicious webpage extract URL features and follow 4,5,6,7,8,9,10 & 11 steps.

4. Checks for the IP address in the URL.

5. It checks for the length of URL.

6. Checking for presence of '@' symbol.

7. Checking for an addition of prefix or suffix.

8. Analyzing for sub-domains.

9. Checking for the trusted user and age if it is through HTTP.

10. Check for whether it is a demand URL.

11. Check for an abnormal URL.

12. Classify the webpage based on above list and show it is a phishing website.

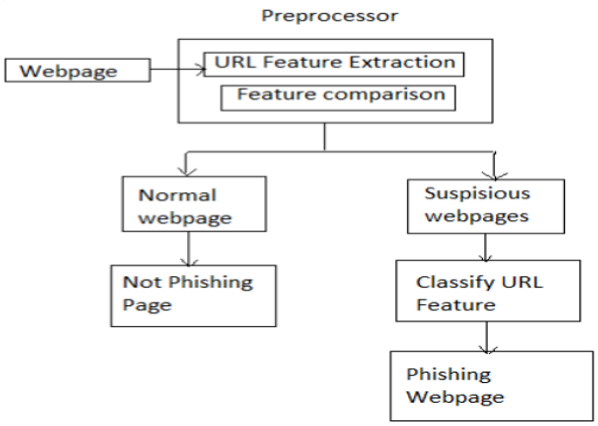

Fig.1. System Architecture

In the classifier module the user accesses the website and then the system analyzes whether the given address is in already saved address or not. If it is in already saved address it is a legitimate website else it is a suspicious URL. After the classifier module classifies the websites, the suspicious URL's are sent to feature extraction module for processing. In the feature extraction module, the suspicious URL features are extracted like length, address, time etc. These features are used to classify whether it is a phishy website or not. In the feature analyzer module several features of URL are compared to detect phishy website.

There are several features distinguish phishing websites from legitimate ones. The features used in our study were explained below.

1. Using IP address: Instead of host name it utilizes IP address as a part of the URL address. It implies clients can nearly make certain somebody is attempting to take their data. This element is a paired element.

Syntax: -http://--+-IP address--+--+---------+--/--*path component*------>

'-host name-'

Eg:- http://63.17.167.23/pc/verification.htm?=https://www.paypal.com/

2. Length of URL: Phishers try to shroud the suspicious part of the URL, which diverts the data presented by the user to a suspicious area. We found that if the URL length is under 54 characters, then the URL is not suspicious, and if the URL length ranges from 54 to 75, then the site is named "suspicious", generally the site is named "Phishing".

Eg:- Genuine website: https://www.h******k.com/

Phishing website: http://paypal.com-webappsuserid29348325limited.active-userid.com/webapps/89980

3. URLs having "@": Phishers try to conceal the suspicious part of the URL. Something that bring about suspicion is the presence of the ""@"" in the URL. The ""@"" image drives the program to overlook everything earlier the ""@"" image and diverts the user to the connection wrote after it.

Eg:- http://olb.westpac.com.au[specialunprintablecharacters]@68.112.112.35:8888/asp/index.html

4. Adding prefixes and additions to URL: Phishers try to mask users by reshaping the URL to resemble the genuine ones. This is done by adding prefix or addition to the genuine URL, and the user may not see any distinction.

Eg:- Genuine URL: http://www.paypal.com.ssl2.us/webscr.php?cmd=LogIn#

Fake URL: https://www.paypal.com/ssl2/us/webscr?cmd=_login-run

5. Sub-domain(s) in URL: In this phishers try to make users believe by including subdomain(s) to the URL, and along these lines, the users may trust that they are operating a genuine site.

Eg:-http://paypal.com.de.cgi-bin.webscr.cmd-login submit.dispatch.sicherkontrolle.su/cgi-bin/

6. Abuse of HTTPs: The presence of HTTPs means a delicate data is exchanged and users get easily confused thinking it is a legitimate site. But phishers can use duplicate HTTPs to fool users.

Eg:http://h.paypal.dechecking.net/de/ID.php?u=LhsdoOKJfsj dsdvg

7. Demand URL: A page comprises of a content and some articles, for example, pictures and recordings. Ordinarily, these articles are stacked on the site page from the same space where the site page exists. On the off chance that the articles are stacked from an area not quite the same as the space wrote in the URL address bar, then the website page is possibly bargained a phishing suspicion. The proportion of the articles stacked from an alternate area distinguishes the worth doled out to this component.

Eg:-
http://www.citibankonline.com/domain/redirect/cbna/global_ nav/myciti.htm?BVP=/&M=S&US&_u=visitor&

BVE=HT%54p%3a%2f%2fkdsass40e.com*20022%2E%64a %2eR%75

8. Abnormal URL: If the site character does not correspond with its record appeared in the WHOIS database, then the site is named phishy. In this if the host name is not in the url then it is definitly phishy.

4. RESULTS AND DISCUSSIONS

Feature	Percentage
IP Address	72%
Lengthy URL	47%
URL having @ symbol	43%
By prefix or addition in URL	87%
Abuse of HTTPs	28%
Demand URL	35%
Abnormal URL	40%

Table.1. Percentage of URL attacks

To prove the proposed system we took 25,320 phishing sites from sites like phish tank etc. And around 2000 legitimate websites from different sources. The URLs taken are made to run through the above system made of URL mining algorithm. After running all url's in the system, observations are made. These observations show that the proposed system detects 98.5% of the Phishing attacks. Table.1 given below classify the percentage of total URL falls under which category after the observations made.

5. CONCLUSION

Phishing means stealing someone information like bank accounts, user-name etc by sending fake e-mails, websites. According to the studies made it proves that users are getting more and more vulnerable to phishing attacks. So, it is necessary for the system proposed to be made necessary for everyone. There are also different ways to detect phishing attacks and they should be researched to make sure phishing is eradicated mostly.

REFERENCES

[1] Aburrous , Hossain MA, Dahal K, Thabtah F. Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Systems with Applications: An International Journal. 2010 December: p. 7913-7921.

[2] Pan , Ding. Anomaly Based Web Phishing Page Detection. In In ACSAC '06: Proceedings of the 22nd Annual Computer Security Applications Conference.; Dec. 2006: IEEE. p. 381-392.

[3] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. Sept.1995;: p. 273 - 297.

[4] Zhang , Hong , Cranor. CANTINA: A Content-Based Approach to Detect Phishing Web Sites. In Proceedings of the 16th World Wide Web Conference; May, 2007.

[5] Manning C, Raghavan , Schütze H. Introduction to Information Retrieval: Cambridge University Press; 2008.

[6] Sadeh N, Tomasic A, Fette I. Learning to detect phishing emails. Proceedings of the 16th international conference on World Wide Web. 2007: p. 649-656.

[7] Bhojane Yogesh, Thakur Yogesh, Apte Omkar ,Bodke Shivam. Intelligent rule-based Phishing Websites Classification. International Journal of Modern Trends in Engineering and Research.2006: p.366-372

[8] Breiman L. I., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and regression trees (cart). Biometrics 40 (3), 358.

[9] Chen, T.-C., Dick, S., Miller, J., May 2010. Detecting visually similar web pages: Application to phishing detection. ACM Transaction on Internet Technology 10 (2), 1–38.

[10] Cherkassky V., 1997. The nature of statistical learning theory . IEEE Transactions on Neural Networks 8 (6), 1564.

[11] Chou, N., Ledesma, R., Teraguchi, Y., Boneh, D., Mitchell, J. C., 2004. Client-side defense against web-based identity theft. In: Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS).

[12] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," IEEE INFOCOM, 2011.

[13] M. Khonji, A. Jones, and Y. Iraqi, "A novel phishing classification based on URL features," IEEE GCC Conf. And Exhibition, 2011.

[14] T. Balamuralikrishna, N. Raghavendrasai and M. Satya Sukumar, "Mitigating online fraud by ant phishing model with URL and image based webpage matching," International Journal of Scientific and Engineering Research (IJSER), March, 2012.