DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4309

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AN EFFICIENT ALGORITHM TO IDENTIFY PHISHING SITES USING URL DOMAIN FEATURES

R.Gowri
M.Phil Research Scholar,
PG & Research Department of Computer Science,
Dr.Ambedkar Government Arts College, Vyasarpadi,
Chennai-600 039

V Karamchand Gandhi
Doctoral Research Scholar
PG & Research Department of Computer Science,
Dr.Ambedkar Government Arts College, Vyasarpadi,
Chennai-600 039

Dr M. Suriakala
Assistant Professor
PG & Research Department of Computer Science,
Dr.Ambedkar Government Arts College, Vyasarpadi,
Chennai-600 039

Abstract: People are using internet for all their day to day activities like shopping, banking, mailing etc. The term phishing is a kind of spoofing website which is used to steal important information. There is a fortune to steal all our personal data by doing something as fraudulent. So there is a need of efficient prevention mechanism to find out the phishing webpage among the legitimate web pages. This paper proposes a study on identifying phishing websites by developing an effective algorithm with domain features of URL. This algorithm is used to overcome the difficulty and complexity of detecting phishing websites which looks exactly like an original websites.

Keywords: Prediction Algorithm, Domain Features, Phishing sites

1. INTRODUCTION

The growth of the phishing websites seems to be astonishing. Even though the web users are aware of phishing attacks, lots of users become victim to these attacks. Numbers of attacks are lunched with the aim of making web users believe that they are communicating with a trusted entity. Phishing is one among them, communications from popular web sites, auction sites, online payment processors are commonly used as a source to lure the unsuspecting public. Phishing websites are mock websites that look similar to legitimate. Only specialists can identify these types of phishing websites immediately. Ying P and Xuhus Ding *[1]* used discrepancies that exist in the website's identity, structural features and HTTP transaction to detect the mock website. It demands neither user expertise nor prior knowledge of the website. The main feature of this approach includes: a)it does not rely on any prior knowledge of the server or user's security expertise; b)the adversary has much less adaptability since the detection is independent of any specific phishing strategy; c)it causes no changes on user's existing navigation behaviour*[2]*.

Anh Le,Athine Markopoulou,university of Calofornia *[3]* used lexical features of the URL to predict the phishing website. Classification accuracy of using lexical feature is compared with accuracy of using automatically selected and hand selected features and compared with additional features.

Many researchers have investigated the detection of phishing websites and the research articles are: Maher Aburrous et al investigated about the e-banking using fuzzy data with two phishing website criteria URL & domain identity and Security and encryption *[4]*. Rajendra Gupta and Piyush Kumar Shukla*[5]* investigated about the novel antiphishing solution and useful to reduce the negative consequences semantic attacks on society by useful security information.

2. PROPOSED APPROACH

Malicious Web sites covers a range of different illicit enterprises which are unsafe to visit, that's why different types of malicious sites allocate various threats to users. If type of this threat is known it will be easy to inspect these types independently and understand their features which will be helpful to track the malicious site and to find out solution against a particular kind of threat*[6]*.

This paper describes the most common features that are used to find the differentiation of legitimate and phishing WebPages based on the URL features. By evaluating all the features, one can determine that the website which resembles the following features considered as phishing. The common features to develop a legitimate website are identified and these features are compared with phishing website features. This is done by the prediction algorithm.

Differences are identified, an algorithm is developed by considering these features to differentiate and identify the legitimate website from the phishing website.

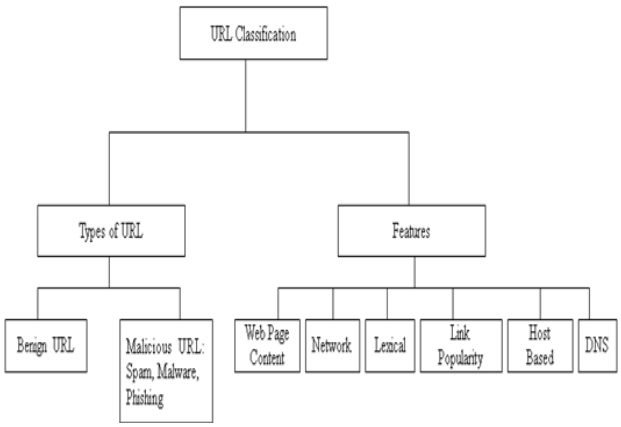


Figure 2.1 URL Classification with types of URL and features

2.1 DOMAIN BASED FEATURE

2.1.1 Domain Age

Most of the phishing website can live for a short period of time but if the website has been for more than a year is good sign of security. This feature can be extracted from the WHOIS database *[7]*. The blacklist may succeed in protecting the users if it works on the domain level not on the URL level that is, add the domain-name to the blacklist not the URL address. However, Rasmussen and Aaron find that 78% of phishing domains were in fact hacked domains, which already serve a legitimate website. Thus, blacklisting those domains will in-turn add the legitimate websites to the blacklist as well *[8]*.

Even though the phishing website has moved from the domain, legitimate websites may be left on blacklists for a long time; causing the reputation of the legitimate website or organisation to be harmed. Some blacklists such as 'Google's Blacklist' need on average seven hours to be updated *[9]*. For this feature, if the domain was created in less than 6 months, it is classified as 'Phishy'; otherwise, the website is considered 'Legitimate'.

2.1.2 DNS Record

If the DNS record is empty or not found then the website is classified as Phishing. DNS record provides information about the domain that is still a live at the URL is not valid any more. For phishing sites, either the claimed identity is not recognised by the WHOIS database or founded cord of the hostname is not found *[10]*.

2.1.3 Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period-of-time thus they may not be recognised by the Alexa database*[11]*.

If the domain has no traffic or is not being recognised by the Alexa database it is classified as Phishy otherwise if the website ranked among the top 100000 it is classified as 'Legitimate' else it is classified as 'Suspicious'.

2.1.4 '@' Symbol

One of the fake signs of fake website is the use of the '@' symbol within the URL Address. This may lead users to neglect all characters before the '@' so attackers can guide users to fake website *[12]*.

2.1.5 HTTPS in URL's Domain

This feature can be used by phishers to deceive users by inserting the 'HTTPS' within the URL's domain *[13]*. For instance, http://https.www.as.co.in.

An Algorithm to identify the phishing WebPages using Domain Based Feature

1. **Input** URL

   Output Legitimate or phishing

2. User requested URL is processed

   i) If age≤6→Phishing //**Age of Domain**

   ii) If no DNS record in the domain→phishing //**DNS Record**

   iii)Webpage rank<100 000→Legitimate //**Website Traffic**

   The ranking>100 000→Suspicious

   Otherwise→Phisy

   iv) If '@' symbol in URL→Phishing // '@'**Symbol**

   v) If HTTPS in URL→Phishing //**HTTPS in URL's Domain**

3. If all condition satisfied and legitimate then

   i) Extract IP Address with domain name

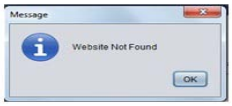
   ii) Source Code will be retrieved

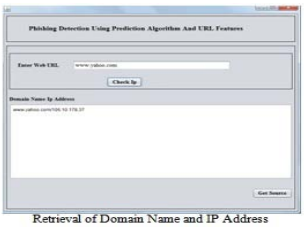
   iii) Attributes are extracted

4. Compute the result

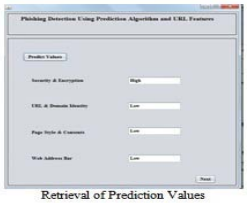
3. IMPLEMENTATION

Java Beans is used to design and execute the entire algorithm. The prediction values can be assigned based on the criteria's according to the phishing indicators. First, the user will enter the URL address to identify the phishing website and an IP address is displayed by using an Internet address. Source code of a particular website is extracted. By using an URL of website, the attributes of website are extracted. Values of a website are also predicted by using prediction algorithm. Finally the result is generated that whether the website is phishing or not. This website feature gives information associated with the current URL and the algorithm can detect based on these values. Phishing websites are identified using predicted values and IP address of a specified website is also retrieved. Retrieving a source code can be done by entering an URL and retrieving time will vary for legitimate and phishing website. Identify the phishing website by analyzing the attributes and predicted value to evaluate the security of the website.





Retrieval of Domain Name and IP Address



Retrieval of Prediction Values

Final result

4. CONCLUSION

Malicious links are well known weapon used by attackers to acquire control of victim systems, which can be utilized to execute cyber crime involving spamming, phishing, denial of service and many more *[14]*. This paper proposes a study on identifying phishing websites by developing an effective algorithm with domain features of URL. Differences are identified, an algorithm is developed by considering these features to differentiate and identify the phishing website. In future, more domain features created day by day in digital era can be added in this algorithm to find phishing websites and warns the user of any possible attack.

5. REFERENCE

[1] Jin-Lee Lee,Dong-Hyun Kim And Chang-Hoon,Lee, "Heuristic-based Approach for Phishing Site detection Using URL Features",CEET-2015.

[2] Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[3] Hao Zhou,jianhua sun and Hao Chen,"Malicious Website Detection and Search Engine Protection ",Journal of Advances in Computer Network,Vol.1,No.3,September 2013

[4] Khonji, Mahmoud,Youssef Iraqi and Andrew Jones."Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

[5] Mohammad, Rami M., Fade Thatch, and Lee McCluskey. "Intelligent Rule-Based Phishing Websites Classification." Information Security, IET 8.3 (2014): 153-160.

[6] Rajendra Gupta and Piyush kkumar Shukla**,** "" Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System",I.J computer Network and Information Security,2015,12,70-77.

[7] Canali,Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World Wide Web,ACM, 2011

[8] Y. Pan and X. Ding, "Anomaly BasedWeb Phishing Page Detection," Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC'06), Computer Society, 2006.

[9] Rami M.Mohammed,fadi thabtah and Lee McCluskey, "Intelligent Rule Based phishing Website Classification",IET Information Security,July 2013

[10] Iuon-Chang Lin and Hung-chieh chuang, "The URL features for phishing by using word Suggestion",International Journal of Advances In Computer Science and Cloud Computing,Vo;ume-3,Nov-2015.

[11] Pan, Y., Ding, X.: 'Anomaly based web phishing page detection'. Proc. 22nd Annual Computer Security Applications Conf. (ACSAC'06), December 2006, , pp. 381–392

[12] Wenyin, Liu, et al. "Discovering phishing target based on semantic link network." Future Generation Computer Systems 26.3 (2010): 381- 388.

[13] Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. "Improved Phishing Detection using Model based feature".CEAS.2008

[14] Mishra, Madhuresh, Anurag Jain Gaurav, and A. Jain. "A Preventive Anti-Phishing Technique using Code word." International Journal of Computer Science and Information Technologies 3.3 (2012): 4248-4250.