A Contemporary Phishing Identification Technique using Prediction Algorithm based on URL Features

R.GOWRI
M.Phil Research Scholar**,**
PG & Research Department of
Computer Science,
Dr.Ambedkar Government Arts College,
Vyasarpadi, Chennai-600 039
gowrirajagopalmca@gmail.com

V KARAMCHAND GANDHI
Doctoral Research Scholar,
PG & Research Department of
Computer Science,
Dr.Ambedkar Government Arts College,
Vyasarpadi, Chennai-600 039
vedhagandhi@gmail.com

Dr M. SURIAKALA
Assistant Professor and Research Advisor
PG & Research Department of
Computer Science,
Dr.Ambedkar Government Arts College,
Vyasarpadi, Chennai-600 039
suryasubash@gmail.com

ABSTRACT:Nowadays people are depending on internet for their personal use as well as business purpose and fraudulent become the problem for the users. The term phishing is a kind of spoofing website which is used to steal important information. This paper identifies an approach of detecting phishing websites by developing a prediction algorithm which is based on URL features and provides an efficient result. This algorithm is used to overcome the difficulty and complexity of detecting phishing websites which looks exactly like an original websites.

Keywords: Phishing Site, Prediction Algorithm, Phishing URL Features

I.INTRODUCTION

Phishing is a term used to describe a malicious individual or group of individuals who scam users. They do so by sending e-mails or creating web pages that are designed to collect an individual's online bank, credit card, or other login information. Because these e-mails and web pages look like legitimate websites, users trust them and enter their personal information. The rapid growth in the number of users has dramatically increased, although internet users are aware of phishing and many fall victim to such as attacks *[1]*. An efficient method is required to identify the phishing websites to protect from stealing important information. Phishing is a web based attack that uses social engineering techniques to exploit internet users and acquire sensitive data. It is vital to minimize online phishing activities due to the fatality of fraudulent problem on all involved stakeholders including online users, banks, businesses and government. As matter fact, preventing phishing activities early is imminent yet a challenging task due to the sophisticated methods used to attack users. There are always innovative ways that created regularly by phishing attackers to confuse the anti-phishing techniques *[2]*. Hence, continues demands are essential to come up with intelligent anti-phishing methods that are based on data mining and machine learning. There are large number of URL features available with a webpage and many of the features are used to identify phishing websites from the original website *[3]*. Phishing scams are typically fraudulent email messages appearing to come from legitimate enterprises (e.g., your university, your Internet service provider, your bank). These messages usually direct you to a spoofed website or otherwise get you to divulge private information (e.g., passphrase, credit card, or other account updates). The perpetrators then use this private information to commit identity theft. Phishing scams are crude social engineering tools designed to induce panic in the reader *[4]*. These scams attempt to trick recipients into responding or clicking immediately, by claiming they will lose something (e.g., email, bank account). Such a claim is always indicative of a phishing scam, as responsible companies and organizations will never take these types of actions via email.

The heuristic based techniques are analyzed and extract features for detecting the websites by using that information *[5]*. Due to the increase of phishing attacks, there is a considerable research focus on phishing detection techniques in recent days. The black list based techniques and the heuristic based techniques are typical phishing detection techniques. In this paper, a prediction algorithm is proposed to detect the phishing websites and proposed technique extracts an attribute of a users requested pages or urls and this technique can detect an IP address of a specified URL and reduce damage caused by phishing attack. This paper uses the heuristic based techniques to analyse and extract phishing site features using the information *[6]*.

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 Mofleh Al-diabat*[7]* investigated about the symmetrical uncertainy(SU) and information gain(IG) to differentiate among features and detect a small set of correlation among feature. Rajendra Gupta and Piyush Kumar Shukla[8] investigated about the novel anti phishing solution and usefull to reduce the negative consequences semantic attacks on society by useful security information.

II.PROPOSED APPROACH

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, the proposed Algorithm can determine that the website which resembles the following features considered as phishing. This is done by the prediction algorithm.

URL FEATURES

ADDRESS BAR BASED FEATURES

IP Address

If the domain part has an IP address in the URL such as “http://125.98.3.145/fake.html”, then it is said to be phishing website i.e. someone is trying to steal our personnal information otherwise it is legitimate site.

Lengthy URL to Hide the Suspicious Part

Often attackers hide the mistrustful portion of the URL to catch a user’s submitted data. They also may edirect the uploaded webpage to a doubtful domain *[9]*. Normally, there is no measure for the URL length but recent studies identified that an accepted URL length is often less than 56 characters.

Phishers can use long URL to hide the doubtful part in the address bar. For example, http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f907l984f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.htm. In this URL, lengthy URL is used to hide some suspicious part to redirect into an suspicious webpage. When the URL contains more than 54 characters then it is considered as an phishing website.

Sub Domain and Multi Sub Domains

Fraudulent websites often have a domain, which has been recently registered, and their lifespan is short. Therefore, when the domain expires in less than a year it can be suspicious *[10]*. If the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.

ABNORMAL BASED FEATURES

Request URL

In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain such as image and videos. If the objects are loaded from the domain other than the typed URL, the webpage is considered as an phishy*[12]*. Irregular URLs: A test to examine whether the current browsed website is inside the WHO-IS database can determine the legitimacy of the website. Here are different objects inside a webpage including text, picture, videos, etc. In cases where the current webpage’s objects are loaded from a server that is different to the URL’s then there is a possibility that this webpage is fake.

Links in <Meta>, <Script> and <Link> tags

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script,*[13]* and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

Abnormal URL

This feature can be extracted from WHOIS database. If the host name is not included in the URL then it is considered as phish *[14]*. Short URL: Sometimes the URL length can be shortened using HTTP Redirect.

III PREDICTION ALGORITHM

Based on the URL features, the phishing attributes are extracted and find out the malicious URL hyperlink.

1**.** Input URL

2. By entering an IP Address, we can get an IP address with an domain name.

2. Extract the source code by using an requested URL.

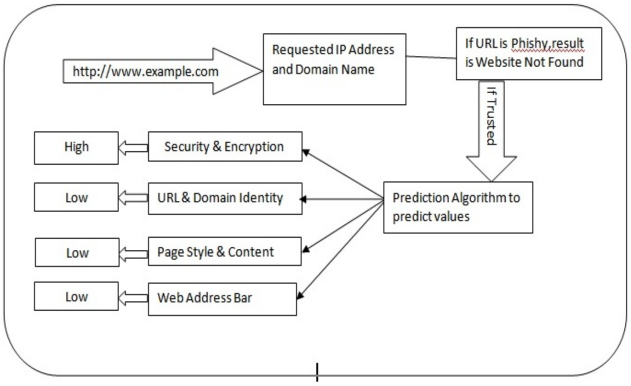
3. Phishing attributes are extracted.

4. If the phishing website URL is entered, produce a result as a Website not found.

Values can be predicted under the following condition:

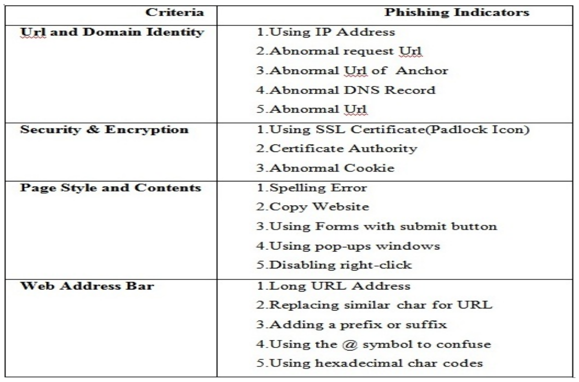
1. URL and Domain Identity = Low
2. Security and Encryption = High
3. Page style and content = Low
4. Web Address Bar = Low

5. Compute the Result.



Process Flow of Proposed Solution

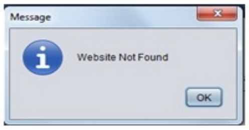
Criteria for phishing website:



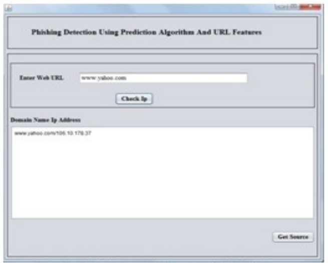
Criteria	Phishing Indicators
Url and Domain Identity	1.Using IP Address 2.Abnormal request Url 3.Abnormal Url of Anchor 4.Abnormal DNS Record 5.Abnormal Url
Security & Encryption	1.Using SSL Certificate(Padlock Icon) 2.Certificate Authority 3.Abnormal Cookie
Page Style and Contents	1.Spelling Error 2.Copy Website 3.Using Forms with submit button 4.Using pop-ups windows 5.Disabling right-click
Web Address Bar	1.Long URL Address 2.Replacing similar char for URL 3.Adding a prefix or suffix 4.Using the @ symbol to confuse 5.Using hexadecimal char codes

IV 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 detected 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. Detecting the phishing website by analyzing the attributes and predicted value to evaluate the security of the website.


Message

Website Not Found

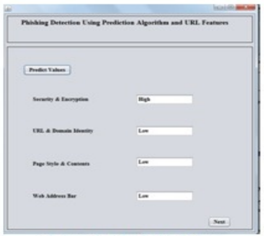

Phishing Detection Using Prediction Algorithm And URL Features

Enter Web URL: www.yahoo.com

Domain Name Ip Address

www.yahoo.com/106.10.178.37

Retrieval of Domain Name and IP Address


Phishing Detection Using Prediction Algorithm and URL Features

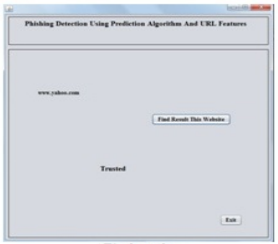
Security & Encryption: High

URL & Domain Identity: Low

Page Style & Contents: Low

Web Address Bar: Low

Retrieval of Prediction Values


Phishing Detection Using Prediction Algorithm And URL Features

www.yahoo.com

Trusted

Final result

V.CONCLUSION

Detecting phishing websites is one of the crucial task in the internet and it is generally difficult to detect and prevent. To detect a fraudulent website, there are many techniques are already found like black list based, heuristic based and page rank etc. This paper proposes prediction algorithm to detect a phishing website based on the phishing indicators like abnormal URL request, length of URL, etc. Phishing websites are detected using predicted value and IP address of 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. In future, more number of URL attributes can be included to extract and it can be implemented with data mining techniques to discover new patterns of phishing URL.

REFERENCES

[1] k Ruth Ramya, K.Priyanka, K.Anusha, Ch.Jyosthna and Y.A Siva Prasad, “An Effective Strategy for Identifying Phishing Websites using Class-Based Approach”International Journal of Scientific & Engineering Research,Volume 2,Issue 12,December-2011

[3] 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.

[4] 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

[5] Pallavi D.Dudhe and Prof.P.L.Ramteke, “Detection of Websites Based on Phishing Websites Characteristics”,International Journal of Innovative Research in Computer and Communication Engineering,Vol 3 Issue 4,April 2015.

[6] V.preethi and G.Velmayil, “Automated Phishing Website Detection Using URL Features and Machine Learning Technique”-Volume 2 Issue 5,sep-oct2016.

[7] Yue Zhang,Jason Hong and lorriecranor “CANTINA:A Content based Approach to detecting Phishng websites”,2007.

[8] Firdous Kausar,Bushra Al-Otaibi,Asma Al-Qadi ,”Hybrid Client side Phishing Website Detection Approach”,International Journal of advanced Computer science and Applications,Vol-5,No.7,2014.

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

[10] 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.

[11] Sunil, A.Naga Venkata and Anjali Sardana. "A pagerank based detection technique for phishing web sites." Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.

[12] Vinnarasi Tharania,R.Sangareswari and M.saleembabu,”Web Phishing detection In Machine Learning using heuristic Image Based Method”,International Journal of Engineering Research and Applications,2012.

[13] 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.

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