Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 3, March 2019, pg.345 – 351

A Review: Phishing Detection using URLs and Hyperlinks Information by Machine Learning Approach

Saad Tayyab; Asad Masood
Department of Computer Science, Riphah International University, Faisalabad, Pakistan
sdgjjr40@gmail.com; asi.mas334@gmail.com

Abstract: In the last few years, many fake websites have developed on the World Wide Web to harm users by stealing their confidential information such as account ID, user name, password, etc. Phishing is the social engineering attacks and currently attacks on mobile devices. That might result in the form of financial loses. In this paper, we described many detection techniques using URL, Hyperlinks features that can be used to differentiate between the defective and non-defective website. There are six main approaches such as heuristic, blacklist, Fuzzy Rule, machine learning, image processing, and CANTINA based approach. It delivers a good consideration of the phishing issue, a present machine learning solution, and future study about Phishing threats by using machine learning Approach.
Keywords: Network Security, Phishing Attacks, Hyperlinks, Social Engineering Websites, Machine Learning Approach

---

1. Introduction:

The term ‘‘phishing’’ first time was used by a hacker’s group in 1996, who stole “America Online (AOL)” accounts information by deceiving using unaware AOL user by input their passwords [1]. Phishing is the fake attempt to attain highly sensitive data like user login information, credit, and ATM cards details by masking a trusty website in a message over the network. Generally getting out by instant messaging or email spoofing it often leads consumers to enter personal data on a fake website, the appearance of which is similar to the genuine site [2]. Which seems to belongs with a trusted and popular organizations or brands, ask about your personal/secret information like bank username, password, contact number, address, banking card details, and so on [3].

More 9,635 phish pages published in the last half of 2008. To overcome this issue, stop visitors to go through these pages is an operative way [4]. The current firewalls, antivirus, and dedicated software do not completely stop the web spoofing attack [5]. Email or message are common sources for carrying these attacks. Phishing has seen as an effective attack for many years, creating a broad range of people. Attackers usually masked as popular social websites, banks, and administrators from IT departments of popular websites like e-commerce sites. These emails may trap users to click on linked URLs to start malware attacks or enter some personal data into a malicious website which has a similar look to an original one [2]. Spam is a wide area of attack that is still not fully assumed. Overall, spam has various methods - chat rooms are focuses on chat spam, blogs are focuses on blog spam [6].

Phishing is one of the supreme thoughtful attacks over the Internet. In which, the end-user inputs their sensitive information like bank account details, password, etc. to the bogus website that appears similar to an original one [7]. Maximum of these websites have a very high graphical resemblance to cheat their sufferers. These types of websites seem the same to the original websites. Sufferers of these websites may input their passwords, bank account details, credit/debit card number, or other personal information on the phishing websites and deliver this information to the owners of these websites [8].

Now maximum people use the internet for many purposes like trading of products, chatting, and mailing purpose. Users spend their maximum time in socializing [9].

The e-commerce, online payment services, and social media are the most suffered by this attack. This attack is achieved by taking benefit of the graphical similarities between the bogus and the original websites [7]. The attacker design and develop a similar webpage to the original webpage. Then transfer the link of this webpage to several hundred users via emails and other ways of communication. Mostly, the bogus email content alarms about fear, threat and alert the user to perform some necessary action. When the user unintentionally updates the private information, the computer-generated criminals got the user's information [7].

The open, nameless and uncontrolled structure of the Internet supports cyber-attacks, which grants serious security weaknesses for networks and for typical users, either for inexperienced and experienced users. While the experience and care of the consumer are significant, it is not feasible totally to stop users from suffering by the phishing attack [10]. There is an incredible increase in suffered consumers as compared to 2016 in 2017. Nearly 4.8 billion peoples use email in 2017 and calculation shows that this number will increase up to 5.6 billion by 2021 [11].

There are numerous tasks in our studies. The main challenge is many phishing Web pages are become lived within 20 hours [19], and URLs change regularly.

In a phishing attack, attacker designed phishing web pages similar to the original webpages to cheat website users to obtain their secret personal and financial information. The attackers then take the important information of the website users, by asking them to insert the secret data on phishing webpage. Ultimately, the attackers can perform financial theft after getting information [13]. The success of detection of phishing websites techniques mostly depends upon the reorganization of phishing websites correctly and within a suitable time [13].

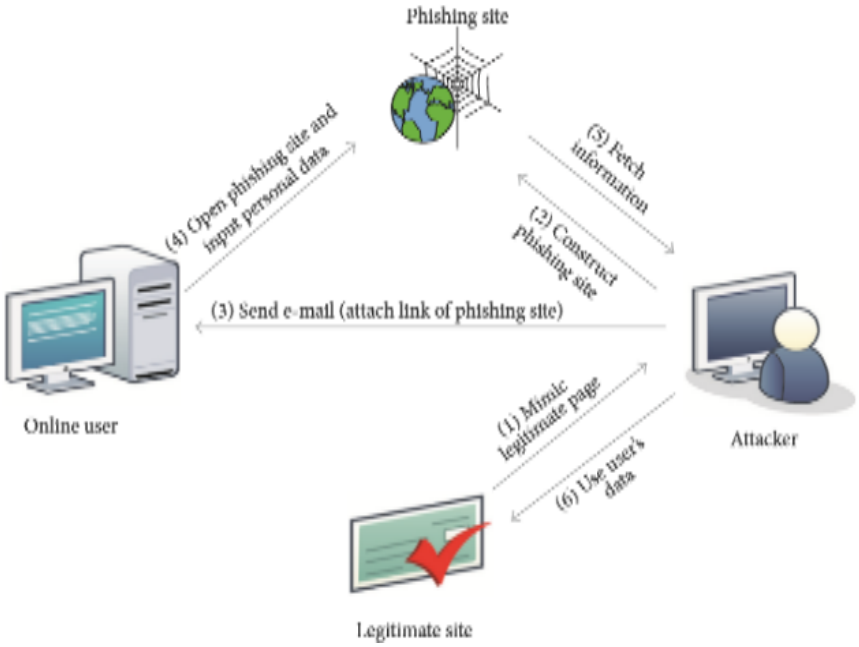

Figure1: phishing mechanism [3]

When the user enters their personal information, which is transmitted, to the attacker. An attacker succeeds in steals the credential of the user [3].

All cyber fraud is a result of a phishing attack that effects the Internet users. According to the APWG phishing report, over 291,096 exclusive phishing websites were spotted between Januarys to June 2017 (APWG H1 2017 Report 2017) [7].

2. Proposed Approach:

Phishing websites is a present issue, however because of its vast effect on the commercial and online selling departments and meanwhile to stop such attacks is a significant step to defending against websites phishing attacks, there are various techniques against this issue and a wide-range collection of related works. In this section, we briefly review already available anti-phishing solutions and associated works.

2.1 Heuristic Approach:

Scientists try to realize the structure of phished websites and senses attacks created on numerous features. Features used in this technique contain a domain name, spelling error, URL, the age of domain, image resources.

2.2 Blacklist Approach:

Blacklists contain URLs whose indicate to those malicious sites. When a browser loads those web pages, it checks blacklist to govern whether the visited URL is in the list or not.[14] 'There have been many works in malicious websites detections efforts. The pioneering effect beings with the blacklist approach[15].'

If it exists, proper security measures should be adopted. Otherwise, the page is considered authentic.

2.3 Machine Learning Approach:

The maximum techniques established to handle the phishing issue depends upon "support vector machine (SVM)". Support Vector Machine is considered a machine learning technique that is significantly used to solve the arrangement issue.

2.4 Fuzzy Rule-Based Approach:

This technique is composed of an arrangement algorithm with few rules drawn from experiments after collecting some unrelated features from a variety of websites as exposed. Those features wide-ranging between three undefined parameters Unsure, Original and Genuine.

2.5 Image Processing Approach:

This technique detected the phishing websites by comparing non-phishy websites with phishy websites depends upon graphical resemblance. This technique breakdown the web pages into chunks based on graphical indications. The graphical resemblance between a phishy and non-phishy site is found by using three metrics: "Chunk level resemblance, layout resemblance, and style resemblance".

2.6 CANTINA Based Approach:

"CANTINA" uses of TF-IDF to detect phishing website. TF-IDF is a popular algorithm to retrieve information which uses for classifying and comparing documents, as well as saving documents from a large mass. In this portion, we first review how TF-IDF works.

- ✓ Give a web page and then evaluate the TF-IDF values of every term of the web page.
- ✓ Create a lexical sign by getting five terms with the highest TF-IDF values.
- ✓ Save this lexical sign to a search engine database, which is Google in our case.

If the domain name of the web page relates the domain name of N top results of the search, we take it as an original web site. Otherwise, a phishing site [14].

Phishing detection techniques are classified into two groups. 1). User training, 2). Trusts on the computer software.

In the user training based technique, Internet consumers are trained to recognize the features of phishing attacks.

In software-based techniques are further divided into graphical resemblance, machine learning, and blacklist approaches. In machine learning technique first trains a cataloging procedure through some features then declares a website as phishing if websites follow the predefined rules to declare, or non-phishing if the websites do not compete with the predefined rule set. Visual parallel based technique links the graphical appearance of the doubtful website and its parallel original website[7]. Our approach decides on the base of 12 features named as "total hyperlink, no hyperlinks, internal hyperlinks, external hyperlinks, null hyperlinks, internal error, external error, internal redirect, external redirect, login form link, external/internal CSS and external/internal favicon". In these features 2, 6, 7, 8, 9, 10 are novel and proposed by the author. 1, 3, 4, 5, 11, 12 Features are taken from another technique [7].

This method is used to develop a new anti-phishing detection URL based tool to stop a phishing attack. Either by screening the phishing URL or by warning the user against the phishing threat [16]. 'Uniform Resource Locators (URLs), sometimes known as "Web links", is the primary source which users use to locate resources over the Internet. Our aim is to find out harmful Web sites from the host-based and lexical features of their URLs [17].

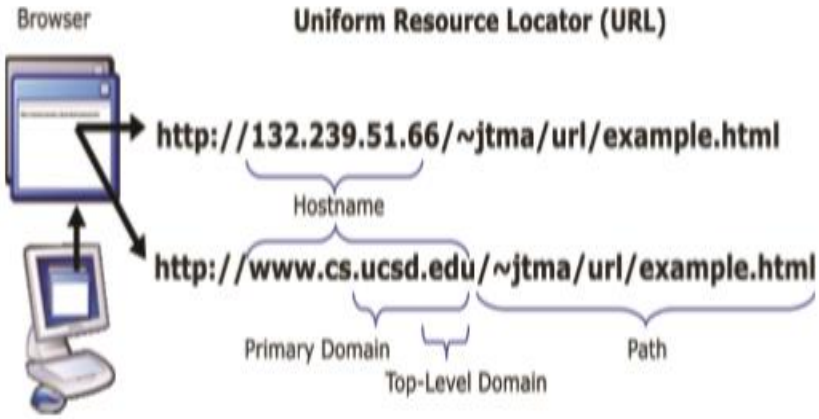

Figure 2:URL & component [17].

3. Problem Overview :

Sometimes URLs are considered as "Web links" and are the main source which the user uses to locate info over the Internet. Our goal is to originate cataloging models that find out phishing websites by analyzing URL's host-based and lexical features. We study various categorizing algorithms in the "Waikato Environment for Knowledge Analysis" (WEKA), "Workbench", and "MATLAB" [18]'.

4. Data Set:

Here datasets used for evaluation is described. For URLs, we use two data sets. 1). "The DMOZ Open Directory Project" [19]. DMOZ is a source in which the editor manually enters the entries. 2). "The random URL selector for Yahoo's directory" [19]. Many times, phishing is done in multiple pages where users click a few links to access the login page with form [20]. Phishing email classifier chooses the best dataset for training that is consist of real emails of both phishing and authentic emails [21].

All classes of "The Microsoft Reputation Service (MRS)" are mined out and kept separate as "Severe, Benign, and Moderate". Multiple classes can be returned for the URL. If a specific URL has two classes, like "Severe" and "Benign" then class with higher threat will prefer. The threat is discussed below.

Extracted Categories: 1. Severe 2. Benign **Hypothesis:** Severe is preferred due to the greater threat Result.

An internal scale decides the category of every URL. **Internal Scale**: either Red or Green or Yellow.

If Result = Severe, Then

Scale = Red

Else If Result = Moderate, Then

Scale = Yellow

Else If Result = Benign, Then

Scale = Green.

This is predicted objective of an attacker to confuse the user by developing phishing page seems possibly similar to genuine [22]. There were 30 features of a webpage that can be used for detection of phishing websites. Those can be mined out using "Python" and used to predict for a new phishing URL [23]. "Beautiful Soup" is Python's library to getting data from " XML and HTML" files [24].

5. Conclusion

Phishing becomes the thoughtful network security issue which affects the financial loss of millions of dollars to users and e-commerce corporations. These attacks can be discovered through an arrangement of customer reportage, bounce checking, image use checking, honeypots, and other methods. We allowed numerous new arrangements to recognize phishing websites. These arrangements were created on the base of the URL of the webpage. We used these arrangements to train "Logistic Regression Classifier", which reached a higher accuracy rate in discovering phishing or non-phishing webpages.

In this paper selection of hyperlink precise structures which are mined from the user. Furthermore, these arrangements are enough to detect a website. The investigational outcomes exposed that the projected technique is very effective in the arrangement of phishing webpages as it has 98.39% accuracy rate. The accuracy of this method may be increased by using some more features. Addition of some more features may improve the arrangement accuracy. However, mining some other features from the third party will improve the cost of the scheme.

References:

[1] H. Thakur, "Available Online at www.ijarcs.info A Survey Paper On Phishing Detection," vol. 7, no. 4, pp. 64–68, 2016.

[2] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," *Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018*, vol. 2018–Janua, pp. 300–301, 2018.

[3] A. K. Jain and B. B. Gupta, "Phishing detection: Analysis of visual similarity-based approaches," *Secur. Commun. Networks*, vol. 2017, no. I, 2017.

[4] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machine learning-based web phishing detection," *3rd Int. Conf. Knowl. Discov. Data Mining, WKDD 2010*, pp. 187–190, 2010.

[5] H. Sampat, M. Saharkar, A. Pandey, and H. Lopes, "Detection of Phishing Website Using Machine Learning," vol. 05, no. 03, pp. 2527–2531, 2018.

[6] A. Bhowmick and S. M. Hazarika, "Machine Learning for E-mail Spam Filtering: Review, Techniques, and Trends," 2016.

[7] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Humaniz. Comput.*, vol. 0, no. 0, pp. 1–14, 2018.

[8] G. Sharma and A. Tiwari, "A Review on Phishing URL Detection using Machine Learning Systems," *Int. J. Digit. Appl. Contemp. Res. Website www.ijdacr.com*, vol. 4, no. 2, 2015.

[9] V. V Satane and A. Dasgupta, "Survey Paper on Phishing Detection : Identification of Malicious URL Using Bayesian Classification on Social Network Sites," vol. 4, no. 4, pp. 2013–2016, 2015.

[10] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, 2019.

[11] S. A. Al-Saaidah, "Detecting Phishing Emails Using Machine Learning Techniques," 2018.

[12] W. Chu, B. B. Zhu, F. Xue, X. Guan, and Z. Cai, "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs," *IEEE Int. Conf. Commun.*, pp. 1990–1994, 2013.

[13] W. Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper

Features Selection," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 72–78, 2017.
[14] P. D. Dudhe and P. L. Ramteke, "A review on phishing detection approaches," *Ijcsmc*, vol. 4, no. 2, pp. 166–170, 2015.
[15] A. Ali Ahmed, "Malicious Website Detection: A Review," *J. Forensic Sci. Crim. Investig.*, vol. 7, no. 3, pp. 1–4, 2018.
[16] R. B. Basnet and T. Doleck, "Towards developing a tool to detect phishing URLs: A machine learning approach," *Proc. - 2015 IEEE Int. Conf. Comput. Intell. Commun. Technol. CICT 2015*, pp. 220–223, 2015.
[17] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to detect malicious URLs," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–24, 2011.
[18] J. James, L. Sandhya, and C. Thomas, "Detection of phishing URLs using machine learning techniques," *2013 Int. Conf. Control Commun. Comput. ICCC 2013*, no. Iccc, pp. 304–309, 2013.
[19] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond Blacklists : Learning to Detect Malicious Web Sites from Suspicious URLs," *Kdd'09*, pp. 1245–1253, 2009.
[20] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, 2011.
[21] A. Yasin and A. Abuhasan, "An Intelligent Classification Model for Phishing Email Detection," *Int. J. Netw. Secure. Its Appl.*, vol. 8, no. 4, pp. 55–72, 2016.
[22] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. Gonzalez, "Classifying phishing URLs using recurrent neural networks," *eCrime Res. Summit, eCrime*, pp. 1–8, 2017.
[23] J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.
[24] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," *Telecommun. Syst.*, vol. 68, no. 4, pp. 687–700, 2018.

