Website Phishing Detection using Heuristic Based Approach

Jaydeep Solanki, Rupesh G. Vaishnav

Computer Department, Darshan Institute of Engineering and Technology, India

---***---

Abstract - *Internet has become a useful part of our regular day to day life as we do almost all of our social and financial activities online. Today every persons are heavily depends on internet and online activities such as online shopping, online Banking, online booking, online Recharge and many more. Phishing is a form of web threat, phisher create the replica of original website and illegally try to get Victim's personal information like user name, password, credit card details, SSN number and use it for own benefit. A Non regular user cannot identify whether website is phished or legitimate. There is no any single solution to stop this fraudulent activity. This paper propose the model which identify the phishing site. System first extract the features which clearly differentiate that whether website are phished or legitimate. Then we apply this features to machine learning techniques it will identify website are phished or legitimate. In this way it will help towards our society.*

Key Words: **Phishing, URL, Classifier, Machine Learning, Phishing Site**

1. Introduction

Phishing is the illegal attempt to acquire confidential information such as credit card details, usernames & password and service security number generally for malicious purpose [1] [5]. The word phishing is originated from word *fishing* because of the way to catch the victims by luring them with fake bait is seems very much similar. Phishing is the act which is more depends on the user then attacker since the user may not be able to identify that the website they have visited is fake or original. This is the point where attackers get advantage to acquire their confidential information like social security number, username, credit card details, passwords, account details etc.

Phishing is typically performed by email spoofing or instant messaging and it usually gives threat message or information updating notification which will redirects victim to the web page to provide their confidential details, The fake website looks very similar to the legitimate website as common user can't identify it until he/ she is regular visitor of that site. It is very normal for the non IT professional person to ignore the threat message and move to the phishing website, Common person can't identify the phishy website with very first view. Even IT professional persons are found to be successfully attacked [1]. As a normal human nature victim do not see or check the details of the page they visit and that is the plus point for the attacker from the victim side.

According to the report released by APWG, 25 September 2015 "Global Survey 1st half of 2014" apple became the worlds most phished brand [4]. 27 May 2015 "Global Survey 2nd half of 2014" top ten companies are targeted constantly and sometimes more than 1,000 time per month [3], Which is very huge amount of attacks and this will lead to the very financial lose And this is where it's not even down, the phishing attacks are rapidly increasing day by day. According to the APWG report, June 23 2014 of Q1 2013 phishing sites leaped by 10.7% and August 29 2014 of Q2 of 2014 128,378 phishing sites were observed which is 2nd highest number of phishing attacks after 164,032 seen in 1st quarter of 2012 [2] [1].

This paper have proposed the website phishing detection model. The rest of the paper is organized as follows: Section 2 provides the literature review of related studies. Section 3 gives detail about relevant and important phishing website features which will be used in the model to distinguish website between legitimate and phishy. Section 4 provides proposed system for detecting phishing sites. Section 5 gives detail about work and its future direction.

2. Related Work

As Cyber Crime is being very popular today and as cyber criminals interest is growing in phishing attacks, because phishing is comparatively more effective and easy tool to use against any common person. Hence Researchers have given many anti-phishing methods. In this section, some of the related works will be discuss.

Basically there are three approaches are employed in website phishing identification. The first one is using list of legitimate or phishy websites so called as Blacklist and Whitelist respectively. In this approach list will be used to identify phishy website. This approach needs already available database of Blacklist and Whitelist which needs to update regularly [1] [2] [7] [10]. The second approach is heuristic-based method in which several features [1] [2] [7] [13] used which are collected from the website and used it to identify it either as phishy or legitimate. In Heuristic-based approach there are two ways so called as content based and non-content based. In content based approach content of the website is used to identify website identity and on the other hand in non-content based approach include URL and host information based features used to detect phishing sites. The third approach is visual based approach in which visual elements of the websites are used to detect identity of the website.

One of the content based anti-phishing technique is CATINA which is very popular and proposed by Zhang et al. In CATINA term frequency-inverse document frequency (TF-IDF) is calculated later few keywords will be extracted from the content of webpage [4].At last lexical signature of it generated and later will be used to search on google and the result will be used for the phishing detection. However, CANTINA fails to track the brand names as keywords and performance will be influenced by the website language also.

Ee Hung Chang et al proposed visual based approach [3]. This method takes the screenshot of the webpage and capture its logo by manually or using logo segmentation process. Then the website logo later will be used for identification. Using image database of google and google image search facility, logo will be match and phishing sites will be detected. This work requires the updated image database.

Many of the other researchers proposed many new features which will gives more power to the detection system. Ahmad Abunadi et al proposed three new features [1] such as Google Page Rank, Google Position and Alexa Rank and compared it with the others using weightage and found to be comparatively more.

3. Important Relevant Features

There are several features which will be helpful to clearly distinguish phishing websites from legitimate ones. Features gives results according to its rule as ‘‘phishy’’, “suspicious” or ‘‘legitimate’’. Based on the studies many new features are proposed by researchers recently but basically there are 27 features which are classified in 6 criteria [1] [2] [6] [7] [9] [1] such as “URL & domain identity”, “Security and encryption”, “Source code and JavaScript”, “Page style and contents”, “Web address bar”, “Social human factor”. Researchers use different numbers of features in their model.

As our model is non content based so mainly used features are “URL & domain name” and “host based”. Which are as follow:

1. Long URL to hide suspicious portion [1] [2] [4] [6] [1]: Attacker use long URL’s to hide the suspecting portion in URL. Scientifically, there is no reliable fixed URL character length. The proposed length of legitimate URLs is 75 characters or less, but there’s no any justification behind this value. For the sake of accuracy assurance an average length is taken. Calculated results showed that if the character length is equals to 54 or more than it, then the URL considered as phishy.

   Rule:
   If {URL Length < 54 → feature = NotLong
   URL Length> = 54 &<= 75 → feature = Suspicious
   Otherwise → feature = VeryLong

2. Number of Dots & Slashes: If URL contains five or more than five slashes then it will consider as a phishing URL [15].

   And if the host portion of the URL has more than 5 or equals to 5 dots [1] [4] [12]. Then webpage is potentially a phishing attack.

   Rule:
   If {No. of Dots >= 5 → feature = Phishy
   otherwise → feature = Legitimate

   If {No. of Slashes >= 5 → feature = Phishy
   otherwise → feature = Legitimate

3. Having @ symbol: phisher generally use @ symbol to trick user. If a URL contains @ then URL classified as a phishy [1] [2] [4] [6] [7] [12]. As Browser might ignore everything prior @ symbol since the real address often follows @ symbol.

   Rule:
   If {URL having @ symbol → feature = True
   Otherwise → feature = False

4. Special Character: IF a URL contains any of this characters such as dash (-), underscore (_), comma (,), and semicolon (;) in it THEN the webpage is considered as phishy [2] [4] [6] [7] [12].

5. HTTP & SSL check: For the security impression of trustworthy and authorised websites use SSL certification secured encryption transaction (https ://) [1] [2] [7] [1]. Generally legitimate websites transfer their confidential information using https: // protocol over internet. Since it is found that even

phishy websites use the https://, we further need to check for the trusted issuer and the SSL certificate age.

Rule:

If {Use http is trusted age >= 2 years → feature = Low

Using http and issuer is not trusted → feature = Moderate

Otherwise → feature = High

6. Request URL: Generally the page content such as video, audio, images etc. are loaded from within the Domain as in address bar [2] [7] [6] [1]. We have to check for the presence of domain in the URL in <Src =>.

Rule:

If {Request URL % < 20% → feature = Legitimate

Request URL %> = 20% &< 50% → feature = Suspicious

Otherwise → feature = phishy

7. IP address: Phisher use IP address in the URL in place of domain name [1] [2] [4] [6] [7] [12] [1]. e.g. 162.54.6.146 and even Sometimes the IP address is converted to its hexadecimal form as in the following link http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html. Here check for the domain part and if it contain hexadecimal value or IP address then it will be considered as phishy.

Rule:

If {IP Address exist in URL → feature = True

Otherwise → feature = False

8. Google Page Rank: Google page rank is service of google to estimate website popularity. It involve all the websites in WWW and very complex calculations in order to link them together. The feature can be extracted from Google page Rank service [1] [15]. If the website has lower rank, no traffic or not in the Google database it is classified as phishy.

Rule:

If {Googlepage rank> 5 → feature = Legitimate

Googlepage rank> = 3 &<5 → feature = Suspicious

Googlepage rank< 3 → feature = phishing

9. Age of Domain: This feature can be extracted from WHOIS database [2] [4] [6] [7] [8] [15]. WHOIS is very popular database which contains information of the websites such as Domain name, Register URL, Registration Expiration Date and many more. The phishing website identification is determined on the basis of the time duration of the domain since it is created.

Rule:

If {Age of Domain >= 2 years → feature = Legitimate

Age of Domain> = 1 &< 2 → feature = Suspicious

Otherwise → feature = Phishy

4. Proposed System

In this section system describe to detect phishy website using Non Content based approach in which website URL will be used as an input in the system.

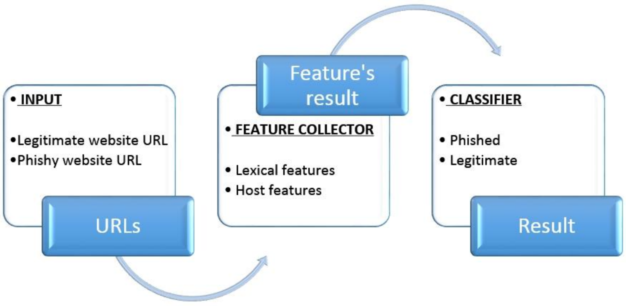

Fig – 1: Proposed System

Step 1: INPUT
Dataset of URL is fed to the model at the initial stage. Which contains 200 Legitimate as well as phishy websites URLs, {collected from the PhishTank and yahoo directory}. Which will be used to Train the Machine Learning Algorithm and Test the performance of it.

Step 2: Feature Collector
It perform Extraction of features from input URL using .NET Script. Such as Lexical features, Host features described in section 3. The result of the every features is extracted using rules and it will be used in later stage. The extracted result of the features is then used to give as an input to the next stage.

Step 3: Classification
This stage use the classifier to finally get result. Classifier is nothing but the machine learning algorithm which is trained to predict the result and perform classification. Since no single classifier is perfect and accurate. The classifiers are chosen mostly because they have been used in the problems similar to ours such as in detecting of spam, phishing emails, phishing websites and malicious URLs etc. system simply wants to use it for final prediction and classification task.

We evaluated Support Vector Machines (SMO) [15] [5] and Decision tree (C4.5 – which is implemented as J48) [15] classifiers to implement it using WEKA (Waikato Environment for Knowledge Analysis) library with their default parameter values. At the end will get result as website is phished or Legitimate.

5. Evaluation

10 Rules (Features) have been applied on input URL's dataset and get value for that features. The output result is in three category Legitimate, Suspicious and Phishy. For final binary output results need to classify it as either Legitimate or Phishy.

Next Support Vector Machine applied on extracted features result and find the value for FP (False Positive) is 5, TP (True Positive) is 120, FN (False Negative) 3 and TN (True Negative) 72. And also have calculated value of F1-measure and Accuracy which are 96.76% and 96% respectively from Precision 0.96 and Recall0.97. The final binary output is analysed using these quality metrics.

Tree is also generated by using decision tree algorithm. Which classify the dataset of legitimate and phishing sites by generation tree structure using features as a decision making point.

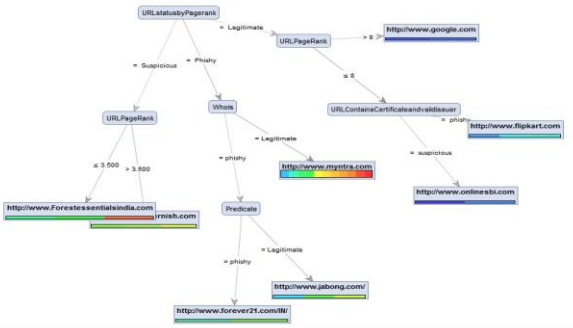

Fig – 2: Decision Tree

6. Conclusion and Future Work

In this paper, heuristic-based phishing detection technique proposed that employs URL-based features. Additionally, classifiers generated through machine learning algorithms and identify the legitimate and phishing websites. System have used SVM which showed accuracy of 96% and very low false-positive rate. The proposed model can reduce damage caused by phishing attacks because it can detect new and temporary phishing sites. System also implemented decision tree algorithm and generated tree for it.

In future work, we will be looking forward for the new features to use and try to improve more accuracy and reduce false positive value of the system. We also look forward to discover new feature with high impact to detect phishing. Also make plugin for the browser which will alert user about phishing website and reduce damage cause with it as much as possible.

REFERENCES

Papers:

[1] D Ahmad Abunadi, OluwatobiAkanbi, Anazida Zainal, "Feature Extraction Process: A Phishing Detection Approach", 3rd International Conference on Intelligent Systems Design and Applications, 2013 IEEE, pp. 331-335, 2013.

[2] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, "An Assessment of Features Related to Phishing Websites using an Automated Technique", International Conference for Internet Technology and Secured Transaction, ICITST 2012. IEEE, London, UK, pp. 492-497. December 2012.

[3] Ee Hung Chang , Kang LengChiew , San Nah Sze and Wei King Tiong, "Phishing Detection via Identification of Website Identity", International Conference on IT convergence and security (ICITCS), 2013 IEEE, pp. 1-4, December 2013.

[4] Y. Zhang, J. I. Hong, and L. F. Cranor., "CANTINA: A content-based approach to detecting phishing web sites", Proceedings of the 16th International Conference on World Wide Web, pages 639–648, 2007.

[5] Mustafa Aydin, Nazife Baykal, "Feature Extraction and Classification Phishing Websites Based on URL", IEEE Conference on Communications and Network Security (CNS), 2015 IEEE, pp.769 – 770, 2015.

[6] Phishing Websites Based on URLPriyanka Singh1, Yogendra P.S. Maravi1, Sanjeev Sharma1, "Phishing Websites Detection through Supervised Learning Networks", International Conference on Computing and Communications Technologies (ICCCT'15), 2015 IEEE, pp 61-65, 2015.

[7] Rami M. Mohammad, FadiThabtah, Lee McCluskey, "Predicting phishing websites based on self-structuring neural network", Springer-Verlag London, Accepted: 10 September 2013, Published online: 21 November 2013. pp. 443-458, November 2013.

[8] Choon Lin, Kang LengChiew, San Nah Sze, "Phishing Website Detection Using URL-Assisted Brand Name Weighting System", IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), December 1-4, 2014 IEEE, pp. 54-59, IEEE, 2014.

[9] Rami Mohammad, "Intelligent Rule based Phishing Websites Classification", IET Information Security Accepted on 21st July 2013, University of Huddersfield, T. L. Mccluskey University of Huddersfield, Vol. 8, Iss. 3, pp. 153–160, July 2013.

[10] Luong AnhTguyen, Nguyen, Ba Lam To, HuuKhuong Nguyen and Minh Hoang Nguyen, "A novel approach for phishing detection using URL-based heuristic", 2014 International Conference on Computing, Management

and Telecommunications (ComManTel), 2014 IEEE, pp. 298 – 303, 2014.

[11] Rami M. Mohammad, FadiThabtah and Lee McCluskey, "Predicting Phishing Websites using Neural Network trained with Back-Propagation", World Congress in Computer Science, Computer Engineering and Applied Computing, Las Vegas, Nevada, USA, pp. 682-686, 2013.

[12] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, "Rule-Based Phishing Attack Detection", International Conference on Security, Institute for Complex Additive Systems Analysis (ICASA), New Mexico Tech., Socorro, NM 87801, USA, 2011.

[13] Luong Anh Tuan Nguyen, HuuKhuong Nguyen, "Developing an efficient fuzzy model for phishing identification", Control Conference (ASCC), 2015 IEEE, pp. 1- 6, 2015.

[14] Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon Lee, "Heuristic-based Approach for Phishing SiteDetection Using URL Features", Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology, Copyright © Institute of Research Engineers and Doctors USA, pp. 131-135, 2015.

Website:

[1] "APWG report of 29 August 2014", [Online], Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf

[2] "APWG report of 23 June 2014", [Online] Available: http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf

[3] "APWG report of 27 May 2014", [Online] Available: www.antiphishing.org/download/document/245/APWG_Global_Phishing_Report_2H_2014.pdf

[4] "APWG report of 25 September 2014", [Online] Available: http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf

[5] http://www.wikipedia.com, Available [Online]

[6] PhishTank, [Online] Available:http://www.phishtank.com

[7] DMOZ, [Online] Available: http://rdf.dmoz.org/rdf/

Books:

[1] Nina Godbole and SunitBelapure, "CYBER SECURITY", Wily India Publishing Pvt. Ltd., New Delhi 2011, pp. 193.

Thesis:

[1] Maher Ragheb Mohammed Abur-rous, PhD Thesis, "PHISHING WEBSITE DETECTION USING INTELLIGENT DATA MINING TECHNIQUES", Submitted for the degree of Doctor of Philosophy, Department of Computing, University of Bradford, 2010.