International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Comparative Study of Detection of Phishing URL

Arun Mishra[1], Gunjan Pareek[2], Brij Kishore[3]
[1]*Student, Department of CSE, Apex Institute of Engineering & Technology, Jaipur, India*
[2,3]*Assistant Professor, Department of CSE, Apex Institute of Engineering & Technology, Jaipur, India*

***Abstract*: In the recent times collecting information has become a trend over the online portals through URL based forms or cookies saving techniques. This paper list out the various techniques of detecting phishing URL using different approaches based on Machine Learning or over the database modeling. A phishing website (sometimes called a "spoofed" site) tries to steal your account password or other confidential information by tricking you into believing you're on a legitimate website. You could even land on a phishing site by mistyping a URL (web address).**

***Keywords*: EML: Extreme Machine Learning, URL Features, Long URLs, Phishing Websites, and Mailers.**

1. Introduction

As networking techniques become increasingly easy to use and efficient, more and more people are using the Internet in their everyday lives. However, even though the Internet can facilitate easy access to information, it can also cause users to lose money easily. Phishing attacks are malicious acts that lure victims to reveal sensitive personal information, such as credit card numbers, bank accounts, and passwords, to fraudulent web pages. According to the latest report from the Anti-Phishing Working Group [1], up to 119,101 unique phishing websites were reported in second quarter 2013 alone; over 74% of these targeted online payment services, and finance-related industries.

In order to protect users from losing money to phishing attacks, many anti-phishing techniques [2] [3] [4] [5] have been proposed to block suspicious web pages. All these techniques have utilized content, non-content, or visual similarity, in identifying web pages. Many of these approaches utilize available data, such as blacklists and search engines. Although many of these methods achieve high levels of accuracy, they have difficulty in keeping up with the fast emergence of phishing web pages.

According to reports [6] [7] from the Anti-Phishing Working Group (APWG), the average number of unique phishing websites detected per day is approximately 1, 500, with each phishing website being short lived (an average of 3.1 days). These statistics indicate that relying on web security experts is not enough to keep up with the emergence of phishing web pages. Some data analysis techniques may also prove helpful in phishing detection. Spam are nothing but the unsolicited bulk emails (UBE) and it's another part is unsolicited commercial email. These spam emails not only consume the user's time but also the energy to recognize the undesired messages, It is wasting the network bandwidth. Content Based filter works on content of emails i.e., text, URLs, main headers like subject for classification purpose. It is the method used to filter spam. The emails include two parts such as Body of the message and Header, Header stores the information about message like from whom it is received, date and time of emails received, sender etc. Now emails ambiguous data is removed by preprocessing then text is extracted.

Many phishing detection techniques have been proposed in the past. These techniques can be grouped into two categories: (1) blacklist-based approaches and (2) heuristic based approaches.

Blacklist-based approaches keep records of all the phishing websites reported by users, or detected by companies. These techniques are widely used in commercial anti phishing tools, such as Internet Explorer, Chrome, and Firefox. They usually achieve a high level of accuracy and are relatively simple to implement. However, since the pre verified phishing list is manually updated, these techniques have difficulty in detecting new phishing websites.

Heuristic-based approaches utilize web content in the detection of phishing behaviors by checking different features, such as visual similarity, or lexical features. The techniques employ visual similarity to calculate the similarity of layout and overall style between potential phishing websites and registered websites, or they compare the similarity of images in pre-stored and trusted websites with those in suspect websites. The visual similarity between two web pages can be based on considering a webpage as a single indivisible entity, or on calculating the signature distances of the images by using Earth Mover's Distance (EMD). These techniques can enhance the accuracy of detection; however, the false hit rate is still unacceptably high.

2. Literature review

In this section we are elaborating the provisos research work which has been proposed to overcome or detect the phishing attacks. The basic techniques to avoid phishing are URL verification, domain check and html contents' scan of web pages referred by email links. Earlier, J. I. H. Zhang proposed

CANTINA [8] to analyze and verify HTML contents of web page refereed by links in emails, domains of URLs found in web pages, also URLs using heuristics approaches. Referring CANTINA, Gupta et. al.[9] has believed in some symbols like "-" that are rarely used in genuine websites. They check for such symbols in domains, URLs and also they check the domain details e.g. age. Also they have used a list of malicious websites against which the scanned URLs are matched. On this basis, web URLs are declared as malicious or genuine ones. But with growing number of websites and domains it will not be easy to update list of malicious or white domains [mahmoud]. Also it is personal to use symbols in websites. Therefore results declared may be unbalanced decision and some genuine URLs can be filtered as malicious.

Justin et. al. [10] detects malicious web sites by extracting properties and features of URLs. The URLs are analyzed and classified thereafter these are matched with a large database which contains filtered malicious URLs. The classification process is performed on real time basis individually by different online classifier which works independently. The database is provided by a mail server which updates it too. Although this work guarantees 99% accuracy to filter URLs but according to Khonji. et. al.[11] large number of entries in data sets can cause performance and resource constraints.

Pradeepthi et. al.[12] has surveyed for classification based methods for detecting phishing URLs and finally proposed that tree based classifier can result with more accuracy. The tree based classifier they define by concluding machine learning and pattern recognition algorithms. This work analyses structure of the URL rather than domain verification and html content mining. However, they again use a data set which is updated at training phase while analyzing URLs.

Gautham et. al.[13] look in html pages, collect the associated direct and indirect links to create a domain set. Also they extract some keywords from the html contents and feed to a search engine which returns another domain set. Concluding a target domain set from the two domain sets they use a third party DNS lookup to check for the legitimacy of the URLs. Using search engine meant for collaborating with again a third party and accuracy of result will depend on results of search engine. Here results are fed by search engine, will be a subject of how the search engine has been designed and defined.

Garera et al. [14] designed 18 hand-selected features to classify phishing web sites, such as page rank information and the period of accessible time. The 18 hand-selected features are very similar to our static characteristic features, but these features still need to query information from the network. Our work extracts the static characteristic features only from the URL string and gives a well representation to distinguish benign and suspicious URLs.

The work by Ma et al. [15] is the most popular research on malicious web site in recent years. They utilize the blacklist, host-based information and lexical information of URLs to build different feature sets. In order to handle large-scale data sets, they use an online learning algorithm and compare the performance with a batch learning algorithm. The results of tests on a data set spanning 100 days are promising. We study their lexical features and adapt them to our own.

Le et al. [16] present a modified version of the method proposed by Ma et al. by using another online learning algorithm. They merge the study from Garera et al. to improve the detection performance.

Thomas et al. [17] design a framework to extract lexical, host-based and page content information as features and implement their idea on a cloud computing platform. This framework is tested on a data set of spam messages gathered from Twitter and e-mail. The result shows the framework achieves high performance at little cost.

Pao et al. [18] proposed using a Kolmogorov complexity-based measure to detect malicious URLs. They adopt a compression method to approximate Kolmogorov complexity, and used the approximation as a significant feature for detection. Unlike their goal of detecting malicious URLs, our work focuses on filtering the large amount of URLs to pick up the most suspicious ones. We use the decision value of the online model to quantify if a URL is sufficiently suspicious to justify downloading the linked content for analysis.

Blum et al. [19] have proposed a similar idea. They try to avoid network queries to reduce processing time, and also choose a confidence-weighted algorithm for lexical information features to detect phishing sites. Different from them, we focus on not only the phishing pages but also all the attacks over the URLs. We also evaluate our framework on a large-scale and extremely imbalanced data set.

The problem in the work by Whittaker et al. [20] is similar to ours in that they too use an imbalanced and large-scale data set. They use the information of host-based, networking, lexical and page content to classify web sites and automatically generate their blacklist. The framework which we proposed does not need to query the information from the network so our system can handle the user query in real time. Several projects have also explored different ways to protect users from malicious URLs.

Li et al. [21] proposed MadTracer which can automatically generate detection rules to detect malicious advertising activities. Invernizzi and Comparetti [22] presented EVILSEED to search malicious web pages more efficiently from an initial seed of known, malicious web page.

Today's internet is suffering from major problem known as Email spam. It annoys users and make financial damage to companies. So far developed techniques to stop spam are filtering methods. Spam emails are UBE also known as junk emails, that are send to many recipients who have not requested or subscribe to this. Spam filter removes spam or un-required messages from email inbox. It also has Phishing URLs which redirects users to phishing websites and seeking personal credentials like username and password for financial purpose.

The existing work by Dhanalakshmi R and Chellapan C, did

implementation on malicious URL detection in Email. Lexical features, page rank, Host information are taken into consideration to classify URLs. Phishtank corpora has been used and Bayesian classification is done to improve the performance of system [23].

Georgios Paliouras et al., have presented learning method to filter spam email. The two machine learning algorithm are considered for anti-spam filtering such as Naïve Bayesian and Memory based learning approach and they are compared concerning performance. So, that in both methods spam filtering accuracy has improved and keyword based filter are used widely for email [24].

Zhan Chuan, LU Xian-liang has given an application for email filtering using a new improved Bayesian filter. They have represented word frequency by vector weights and word entropy is used for attribute selection then formula is derived which improves the performance apparently [25].

Vikas P. Deshpande et al., has presented an efficient method of naïve Bayesian which blocks all spam emails without blocking legitimate emails. To derive solution on this problem, they considered statistical classifier such as naïve Bayesian anti-spam filter and content based spam filter which are adaptive in nature [26].

Sheng et al., have shown that phishing websites are hacked as soon as they are identified as phishing campaigns have two hours of average life. So to block and identify such phishing URLs they have extracted features like suspicious characters, number of dots, IP address, hexadecimal character [27].

Pawan et al., discovered malicious URLs by enhancing blacklisting. One conflict with this method is that their updation process is fast so they failed to identify phishing URLs in early hours of a phishing attack [28].

Maher Abburrous et al., endeavor for a survey to recognize the essential features which can develop accuracy and precision for malicious URLs detection [29].

Congfu Xu et al, did a feature extraction on Base64 encoding of image with n-gram technique. A SVM needs to be trained for efficiently detecting spam images from legitimate images. Its seen from experiment that It has improved the performance in terms of Accuracy, Precision and Recall [30].

R. Malathi et al., has given a new spam detection method by employing Text Categorization, using Supervised Learning with Bayesian Neural Network which uses Rule based heuristic approach and statistical analysis tests to identify “Spam” [31].

Sadeghian A. et al, had presented spam detection based on interval type-2 fuzzy sets. This system gives user more control on categories of spam and permits the personalization of the spam filter [32].

CANTINA+ classifies phishing URLs and the feature set is more exhaustive and obtained classification accuracy of 92.3%. There exist various related researches and case studies conducted on analyzing the feature set required to reduce the exhaustiveness and time consumption [33].

3. Conclusion

After reading out the several papers and postulates presented by them it is evident that detecting of phishing URLs needs constant updates of algorithms and more prominently the database which is being used as the reference for checking out the phishing URLs, hence the inclusion of machine learning is definite but with this the algorithms needs to update their database with each iteration and have to add it to their preferences and score matrices to be able to predict the best result with improving accuracy after every iteration.

References

[1] The apwg website, 2013.
[2] S. Afroz and R. Greenstadt. Phishzoo: Detecting phishing websites by looking at them. In 2011 Fifth IEEE International Conference on Semantic Computing, pages 368–375, 2011.
[3] J. Ma, L. K. Saul, S. Savage, and G. M.Voelker. Identifying suspicious urls an application of large-scale online learning. In ICML ’09 Proceedings of the 26th Annual International Conference on Machine Learning, pages 681–688, 2009.
[4] N. Chou, R. Ledesma, Y. Teraguchi, J. C. Mitchell, et al. Client-side defense against web-based identity theft. In NDSS, 2004.
[5] Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: A content based approach to detecting phishing web sites. InWWW’07 Proceedings of the 16th international conference on World Wide Web, pages 639–648, 2007.
[6] APWG et al. Phishing activity trends: Report for the month of January, 2008, 2008.
[7] APWG. Phishing activity trends report 4th quarter 2012. APWG Industry Advisory, 2013.
[8] J. I. H. Zhang, Yue and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
[9] Ma, Justin et. al., "Learning to detect malicious urIs," in ACM Transactions on Intelligent Systems and Technology (TISn, 2011.
[10] K. V. Pradeepthi and A. Kannan, "Performance study of classification techniques for phishing url detection," in Advanced Computing (ICoAC). lEE, 2014.
[11] Y. I. Khonji, Majid and A. Jones, "Phishing detection: a literature survey." in Communications Surveys & Tutorials. IEEE, 2013, pp. 2091-2121.
[12] I. K. Ramesh, Gowtham and K. S. S. Kurnar, "An efficacious method for detecting phishing webpages through target domain identification," Decision Support Systems, 2014.
[13] J. J. Gupta, Anjali and K. Thakker, "Content based approach for detection of phishing sites," in International Research Journal of Engineering and Technology, vol. 2. IRJET, 2015.
[14] S. Garera, N. Provos, M. Chew, and A. D. Rubin. A framework for detection and measurement of phishing attacks. In Proceedings of the 2007 ACM workshop on Recurring malcode, WORM ’07, pages 1–8, New York, NY, USA, 2007. ACM.
[15] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Identifying suspicious urls: an application of large-scale online learning. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, pages 681–688, New York, NY, USA, 2009. ACM.
[16] A. Le, A. Markopoulou, and M. Faloutsos. Phishdef: Url names say it all. In INFOCOM, pages 191–195. IEEE, 2011.
[17] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. In Proceedings of the 2011 IEEE Symposium on Security and Privacy, SP ’11, pages 447– 462, Washington, DC, USA, 2011.
[18] IEEE Computer Society. H. K. Pao, Y. L. Chou, , and Y. J. Lee. Malicious url detection based on kolmogorov complexity estimation. In The 2012 IEEE/WIC/ACM International Conference on Web Intelligence, Dec. 2012.
[19] A. Blum, B. Wardman, T. Solorio, and G. Warner. Lexical feature based phishing url detection using online learning. In Proceedings of the 3rd ACM workshop on Artificial intelligence and security, AISec ’10, pages 54–60, New York, NY, USA, 2010. ACM.

[20] C. Whittaker, B. Ryner, and M. Nazif. Large-scale automatic classification of phishing pages. In NDSS. The Internet Society, 2010.
[21] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In T. Yu, G. Danezis, and V. D. Gligor, editors, ACM Conference on Computer and Communications Security, pages 674–686. ACM, 2012.
[22] L. Invernizzi and P. M. Comparetti. Evilseed: A guided approach to finding malicious web pages. In IEEE Symposium on Security and Privacy, pages 428–442. IEEE Computer Society, 2012.
[23] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", in AASRI Conference on Intelligent Systems and Control, vol. 4, pg. 125–131, 2013.
[24] G. Paliouras et. al., "An Evaluation of Naive Bayesian Anti-Spam Filtering", in Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, pages 9–17, 2000.
[25] Zhan Chuan et al., "An Improved Bayesian with Application to Anti-Spam Email", in Journal of Electronic Science and Technology of China, Vol.3 No.1, Mar. 2005.
[26] Vikas P. Deshpande and Robert F. Erbacher, "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques", in Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007.
[27] Sheng, S. et al., "An empirical analysis of phishing blacklists", in Proceedings of the CEAS'09, 2009.
[28] Pawan Prakash et al., "PhishNet:Predictive Blacklisting to Detect Phishing Attacks", in Proceedings of the IEEE Infocom, pp.1-5, 2010.
[29] Maher Aburrous et al., "Experimental Case Studies for Investigating EBanking Phishing Techniques and Attack Strategies", Cognitive Computing, Vol. 2, pp. 242-253, 2010.
[30] Congfu Xu et al., "An approach to image spam filtering based on base64 encoding and N-Gram feature extraction", in IEEE International Conference on Tools with Artificial Intelligence.
[31] R. Malathi, "Email Spam Filter using Supervised Learning with Bayesian Neural Network", Computer Science, H.H. The Rajah's College, Pudukkottai-622 001, Tamil Nadu, India, Int J Engg Techsci Vol 2(1),89-100, 2011.
[32] Sadeghian, A and Ariaeinejad, R., "Spam detection system: A new approach based on interval type-2 fuzzy sets", in IEEE CCECE -000379, 2011.
[33] Xiang, G. et al., "CANTINA+: A feature-rich machine learning framework for detecting phishing Web sites". in ACM Trans. Inf. Syst. Secur. Vol.14, No.2, pp.1-21, 2011.