IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 5, May 2021

DOI: 10.17148/IARJSET.2021.8551

Real-Time Phishing Detection – A Literature Survey

Uthkarsh Sanjay [1], Akshay G R [1], Pushkar Anand [1], Adith A Danthi [1], Ravi P [2]

Department of Computer Science and Engineering

Vidyavardhaka College of Engineering, Mysuru, Karnataka India[1, 2]

Abstract-Phishing is a serious cybercrime that is affecting thousands of people on the internet. Nowadays Cyber Attacks are reaching to end users , taking advantage of the weakest security element. People have come out with many solutions to lower the negative effects. Thus in our paper, we are reviewing many proposed techniques of phishing, which includes detection , correction , prevention which is integral to detect phishing. In this literature survey, we are comparing various methodologies used to detect the phishing attack.

Key Words : Phishing, TF-IDF, Random Forest, KNN, Fuzzy Model.

I INTRODUTION

Phishing attacks target the weak areas that are present due to the human errors at the cost of valid information like, credit card info, employment details, social security, bank information. These fraudulent use false websites, electronic mails made up to deceive users involved in confidential financial data transactions by gathering credentials. The Innocent user's belief in the information they find on the internet and phishers operate injection attacks by means of email/website/ url redirection.

Phishing tricks are increasing , one of them is to project a login screen wherein it lets them replicate the identical website. The Phishers dispatch an email which has Hyperlink which is redirected to a pristine website, which claims to be legal. But may request for valid account information like official websites. So, it is clear that phishers use tricky mechanisms to tempt the users using suspicious URLs, email, iframe, suspicious script,images.

The General Phishing-Detection increases accuracy by deploying feature selection algorithm. By selecting many features of the dataset the algorithm chooses ones that are important in predicting the outcome. Irrelevant features do not impact the accuracy of the system. Further, the system is trained with the help of Ensemble Learning. Using multiple models , while performing predictions the result is unbiased hence, it is concluded that the outcome from all the models are considered to depict the majority. Example, if most of the models alarm that a website is phishing, then the conclusion drawn from the ensemble shows the website is in fact phished.

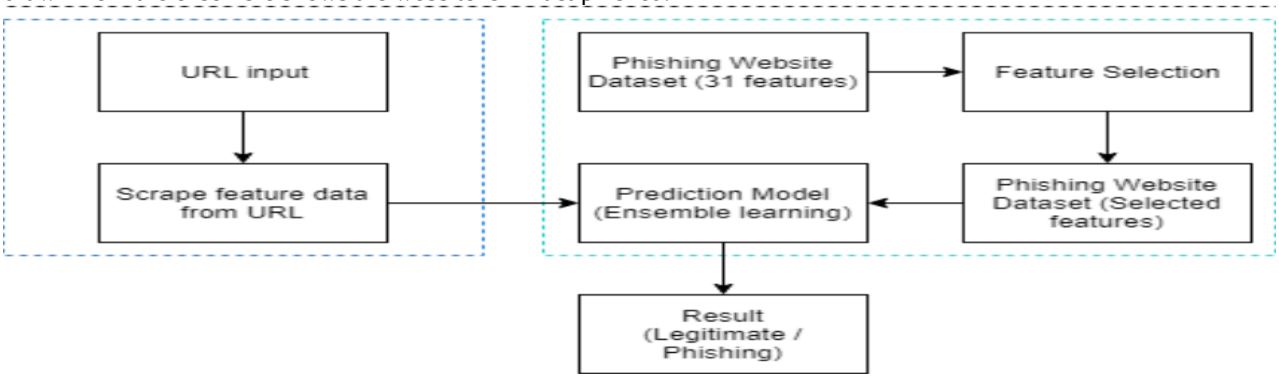

Fig 1: General Phishing Detection Model

II LITERATURE SURVEY

In this paper the methods used are One Class SVM, Linear SVC classifier, K-Nearest Neighbour, Decision tree classifier, Random Forest Classifier split into 2 stages Creation and Prediction and the result stated that Random Forest has proven high precision of 96.87% when compared to other algorithms[1]. In this paper the methodology used is Random Forest Classifier on a Public Dataset and the result observed is that, this method fared better than the others with the highest

This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 5, May 2021

DOI: 10.17148/IARJSET.2021.8551

accuracy of 97.36%[2]. In this paper the methodology used is TF-IDF weights to words similar to the hostname, path and filename URLs. These were then put through a WHOIS search to check if there was any discrepancy with the original and the selected domain name and the result obtained is A phishing website can be differentiated if the query domain name and owner domain name differ[3]. In this paper MFPD is achieved by Using CNN to extract local correlation features from UR.LSTM network sequential dependency from character sequence and softmax is used to classify selected features and the result observed that the MFPD approach proves to be productive with more accuracy, less false positive rate as well as high noticing speed[4]. The methods used in this paper are Fuzzification, Rule Evaluation, Aggregation of the rule outputs, Defuzzification & the result observed is The fuzzy website phishing system showcased the importance of the phishing website criteria by layer one, and proved that the website could be certainly phishy, when rest of the characteristics are evident and true[5]. This paper uses Google Image Database to find out the recognition of the segmented website Logo. Context based image retrieval mechanism is used to match the identity in Google Image Search engine and accuracy obtained is Detection accuracy is increased up to 93% on Google image database[6]. The methods used in this paper is PhishNet: Predictive Blacklisting, DNS-Based Blacklist, Google Safe Browsing API, Automated single White-List and accuracy observed is Blacklists are frequently updated lists of phishing URLs, protocols which are previously detected[7]. This methods works by checking two website page's codes for real and false websites and comparing the safety percentages between them by extracting some phishing characteristics out of the W3C standards and the observation is that High percentage shows a secured website , as well as the others specify the website is certainly to be a phished one[8]. This method works by comparing the Similarity between two web pages is found by contrasting the content of the two websites and the precision obtained is This method detects the phished website pages having accuracy of 0.96 and false-rates not exceeding 0.105[9]. The method in this paper works as follows the keywords in the URL are converted into normal images and then present their image signatures with features comprising of major color category along with its centroid coordinate to find the similarity of two Web pages and this method detects phishing Only if it is visually similar it does not take the code into consideration[10]. There are different Classification Methods implemented like Linear Discriminant, Na¨ıve Bayesian, K-Nearest Algorithm in this paper and this approach has a true accuracy of 85%-95% and its false rates lies between 0.43%-12%[11]. Once the target domain of the suspicious web page is identified,a third party DNS lookup is performed and the two IP's are checked for similarity in this paper and The results show that this system has 99.85% domains correctly identified[12]. In this paper the steps used to check are, Abnormal Anchors,Abnormal Server Form Handler, Abnormal Request URL, Abnormal cookie, Abnormal certificate, Abnormal URL, Abnormal DNS record, in SSL and has resulted that false-positive rate as well as miss-rate are extremely low[13]. In this paper TF-IDF Algorithm is used to detect phishing and Robust Hyperlinks is applied to search the owner of those brands and the observed result is that the TF-IDF method can detect 97% of the fake sites having 6% false-positive[14]. The approach in this paper is Heuristic-based that examines more than one characteristic of a site to identify phishing and the test resulted in 98% of phishing detection rate[15]. This paper is using different Features like Type Based, Domain Based, Page Based, Word Based Features and we observed that approximately 777 unique phishing pages were found on a single day and 8.24% of the users who view phishing pages can be categorized as potential phishing victims[16]. The method used in this paper works in 4 steps : Retrieve Potential Phishing Sites,Send URL to Workers, Worker Evaluates Potential Phishing Site, Task Manager Aggregates Results and out of all the tools available, IE7 was the only tool that could correctly recognize 60% of phishing URLs, yet miss classified 25% of the APWG made-up URLs and 32% of the phishtank.com URLs[17]. In this paper various normal data proportions for test and training converge to a single avg value, and expect to get more objective results and we got to know that here SVM is identified to be finer to NN in detection; regarding the not-true alarm rate and in accuracy for Probe, Dos and U2R and R2Lattacks, in terms of accuracy only NN could outperform the SVM[18]. In this Paper, the steps to stay safe from phished websites are explained and made sure that everyone is aware of basic guidelines to follow before giving our personal details to the fake website and as a result it is expected the users to make sure to doubly check the website before giving our personal and important information to the fake websites[19].

IIIConclusion

Phishing is a serious cybercrime that is affecting thousands of people on the internet. It has increased over the years as more people move online. We need a reliable solution to prevent these cyber criminals from exploiting individuals of their savings.

After going through the above listed papers on real-time phishing detection, it is clear that there are a lot of different approaches to solve this problem with each approach having its own advantages and potential disadvantages.

REFERENCES

1. H L, Gururaj & BoreGowda, Goutham. (2020). Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology. 5. 1-14. 10.1080/23742917.2020.1813396.

This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 5, May 2021

DOI: 10.17148/IARJSET.2021.8551

2. Subasi, A., Molah, E., Almkallawi, F., & Chaudhery, T. (2017). Intelligent phishing website detection using random forest classifier. *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 1-5.
3. Tan, Choon Lin & Chiew, Kang Leng & Sze, San. (2015). Phishing website detection using URL-assisted brand name weighting system. 2014 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2014. 54-59. 10.1109/ISPACS.2014.7024424.
4. P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.
5. M. Aburrous, M. A. Hossain, F. Thabatah and K. Dahal, "Intelligent Phishing Website Detection System using Fuzzy Techniques," 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, 2008, pp. 1-6, doi: 10.1109/ICTTA.2008.4530019.
6. A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016, pp. 1-6, doi: 10.1109/IEMCON.2016.7746247.
7. E. H. Chang, K. L. Chiew, S. N. Sze and W. K. Tiong, "Phishing Detection via Identification of Website Identity," 2013 International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1-4, doi: 10.1109/ICITCS.2013.6717870.
8. Efe-Odenema, Omejevwe & Jaiswal, Jitendra. (2020). Issue 6 www.jetir.org (ISSN-2349-5162).
9. Mona Ghotaish Alkhozae, Omar Abdullah Batarfi, "Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code". Volume 1 No. 6, October 2011.
10. A. Y. Fu, L. Wenyin and X. Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)," in IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301-311, Oct.-Dec. 2006, doi: 10.1109/TDSC.2006.50.
11. J. H. Huh and H. Kim, "Phishing detection with popular search engines: Simple and effective," in Proceedings of the 4th Canada- France MITACS Conference on Foundations and Practice of Security, ser. FPS'11. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 194–207. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-27901-0 15.
12. G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," Decision Support Systems, vol. 61, no. 0, pp. 12 –22, 2014. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S0167923614000037
13. Pan, Y. and X. Ding. "Anomaly Based Web Phishing Page Detection." *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)* (2006): 381-392.
14. Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. 2007. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th international conference on World Wide Web* (*WWW '07*). Association for Computing Machinery, New York, NY, USA, 639–648. DOI:https://doi.org/10.1145/1242572.1242659.
15. Dunlop, M., Groat, S., & Shelly, D. (2010, May). Goldphish:Using images for content-based phishing analysis.In Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on (pp. 123-128). IEEE
16. Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. 2007. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode(WORM '07)*. Association for Computing Machinery, New York, NY, USA, 1–8. DOI:https://doi.org/10.1145/1314389.1314391
17. Zhang, Yue & Egelman, Serge & Cranor, Lorrie & Hong, Jason. (2007). Phinding phish: Evaluating anti-phishing tools.
18. S.Al-Sharafat, W. "Development of Genetic-based Machine Learning for Network Intrusion Detection (GBML-NID)". World Academy of Science, Engineering and Technology, Open Science Index 31, International Journal of Computer and Information Engineering, (2009), 3(7), 1677 - 1681.
19. Anti-Phishing Working Group Phishing, (2014). Anti-Phishing Working Group Phishing Trends Report. [Online]Available at: https://apwg.org/ [Accessed 30 Mar. 2015].

This work is licensed under a Creative Commons Attribution 4.0 International License