DHANALAKSHMI SRINIVASAN
COLLEGE OF ENGINEERING
Coimbatore-641 105
(Approved by AICTE, New Delhi || Affiliated to Anna University, Chennai)

WEB PHISHING DETECTION

(Domain: Applied Data Science)

Literature Survey

TEAM MEMBERS :

1. Ashiq Mohammed M – 721919104019 – IV CSE

2. Pradeesh E – 721919104040 – IV CSE

3. Jeevanantham M – 721919104026 – IV CSE

4. Anandhu B S – 721919104008 – IV CSE

A Literature Survey on WEB PHISHING DETECTION :

In this section, it was discussed some of the techniques which based on list, rule, visual similarity, and machine learning.

A. List Based Phishing Detection Systems

These systems use two lists to classify phishing and non-phishing websites. These are called whitelist and blacklist. The whitelist contains safe and legitimate websites, while the blacklist includes websites classified as phishing.

In 'Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management - DIM 08, pp. 51–60, 2008.', researchers used the whitelist to identify phishing sites. In the study, access to websites takes place only on the condition that the URL is in the whitelist. Another method is the blacklist approach. In the literature, apart from applications such as Google Safe Browsing API, PhishNet, there are also some studies using blacklists like 'M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840–843, 2008'. In blacklist-based systems, the URL is checked from the list and access to the URL if it is not included in the list. The biggest disadvantage of these systems is that the small change in the URL prevents matching in the list. Additionally, the newest attacks, which are named zero-day attacks, cannot be catches with these type protection systems.

B. Rule-Based Phishing Detection Systems

In these systems, features are obtained based on relational rule mining. The rules are estimated to emphasize features that are more common in phishing URLs 'M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013'. In studies using this type of system, it is aimed to use effective features more actively in the classification. In these systems, a set of rules are determined. Thus, the system gives a higher accuracy rate when trained with these rules.

In this context, like CANTINA study '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 - WWW 07, pp. 639-648, 2007.', the Term Frequency - Inverse Document

Frequency (TF-IDF) and rules were used to detect phishing attacks. In addition, in similar studies, models were created by using some features and rules.

C. Visual Similarity-Based Phishing Detection Systems

These systems are based on visual similarity comparison of the web pages. Phishing and non-phishing sites are classified by taking a server-side view of them. These two data are compared with image processing techniques. Fake websites are often designed very close to the original ones. But visually, there are minor differences between them. It is easier to notice these differences, which users cannot easily notice, with image processing techniques. According to the similarity obtained, it is decided whether the website is phishing or not. In the literature, as in the study 'L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide Web - WWW 05, pp. 1060-1061, 2005.', there are studies, which detect the differences based on basic similarities.

D. Machine Learning Based Phishing Detection Systems

The detection of the phishing website in Machine Learning Based Phishing Detection Systems is based on the classification of the specified features by using some artificial intelligence techniques. Features are created by collecting in different categories such as URL, domain name, website features or website content etc. Due to the dynamic structure, especially for the detection of the anomaly in the web sites, it has high popularity on the security of the users.

In the literature, there are some works on this type of detection mechanism. The previously mentioned CANTINA project '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 - WWW 07, pp. 639-648, 2007' was also done by using the machine learning method. According to TF-IDF and heuristic approaches, they detected a 90% accuracy rate. Researchers developed a phishing protection system called PhishWHO applied with three steps in 'C. L. Tan, K. L. Chiew, K. Wong, and S. N. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder," Decision Support Systems, vol. 88, pp. 18–27, 2016.' to determine if the website is legitimate. According to their 3-Tier Identity Matching System, they detected a 96.10% accuracy rate. In 'A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," 2011 Proceedings IEEE INFOCOM, pp. 191-195, 2011.', phishing websites are defined by classifying them with URL attributes such as length, number of special characters, directory, domain name, and file name. The title and priority order of the incoming email is discussed in 'R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," Journal of Network and Computer Applications,

vol. 36, no. 1, pp. 324–335, 2013.'. In 'S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing url detection using association rule mining," Human-centric Computing and Information Sciences, vol. 6, no. 1, Oct. 2016.', URL-based features are used together with features related to transport layer security (Length, slash number, point number and location). They detected the accuracy rate 93% by using the rules obtained by apriori algorithm. In 'M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," Soft Computing, vol. 23, no. 12, pp. 4315–4327, 2018.', a nonlinear regression strategy is used to determine whether a website is phishing. The system was operated by using the harmony search and Support Vector Machine (SVM) methods. They used 11055 webpages and 20 features. Features were selected by using the decision tree method instead of the wrapper. They detected the accuracy rate 92.80% by using nonlinear regression based on HS led. In another study 'E. Buber, B. Diri, and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), pp. 337-342, 2017.', a phishing detection system with 209 word-vector features and 17 NLP based features was proposed. The Random Forest, SMO, and Naïve Bayes algorithms were compared, and the best result was obtained with the Random Forest algorithm in the hybrid approach with an accuracy rate of 89.9%.

In the system proposed in 'E. Buber, B. Diri, and O. K. Sahingoz, "NLP Based Phishing Attack Detection from URLs," Advances in Intelligent Systems and Computing Intelligent Systems Design and Applications, pp. 608–618, 2018.', the number of NLP vectors was increased, and three different machine learning algorithms were compared according to their accuracy values. The Random Forest, SMO, and Naïve Bayes algorithms were compared, and the best result was obtained with the Random Forest algorithm in the hybrid approach with an accuracy rate of 97.2%. Researchers implemented a phishing detection system in 'R. M. Mohammad, F. Thabtah, and L. Mccluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2013.' by using adaptive self-configuring neural networks for classification. In the study, 17 different features are used, which are also used third-party services. Therefore, it was stated that much more time is needed in real-life implementation. In 'A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2017.', to distinguish phishing websites from legitimate ones, a machine learning method was used with 19 features from the URL and Source code, which do not depend on any third party. The results showed that, by using this system, a 99.09% accuracy rate was calculated.

In ‘F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han & J. Wang, “The application of a novel neural network in the detection of phishing websites,” Journal of Ambient Intelligence and Humanized Computing, pp 1-15, 2018.’, the neural network-based classification method is proposed for the detection of phishing websites using the Monte Carlo algorithm and risk reduction principle. ‘G. Karatas and O. K. Sahingoz, "Neural network based intrusion detection systems with different training functions," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-6, doi: 10.1109/ISDFS.2018.8355327.’ focused on the effect of the training functions on neural network to increase the efficiency of the proposals. In ‘S. Smadi, N. Aslam, and L. Zhang, “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning,” Decision Support Systems, vol. 107, pp. 88–102, 2018.’, four different categories are specified: e-mail headers, URLs in the content, HTML content, and main text. The classification was made in machine learning by using 50 features in these categories. The results demonstrated a 98.6% accuracy rate. In ‘R. S. Rao and A. R. Pais, “Detection of phishing websites using an efficient feature-based machine learning framework,” Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, Jun. 2018.’, machine learning algorithms are compared with the features extracted from URL, source code, and third-party services. Principal component analysis Random Forest performed the accuracy of 99.55% with also detecting zero-day phishing attacks. In an NLP-based study ‘T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 300-301, 2018.’, the text of e-mails was analysed and classified. In ‘R. S. Rao, T. Vaishnavi, and A. R. Pais, “CatchPhish: detection of phishing websites by inspecting URLs,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 813–825, Oct. 2019.’, classification was made using TF-IDF, hand-crafted features, and both, along with 35 features. In the study, the phishing attack detection rates were compared by using 6 different algorithms. The best result was obtained in the Random Forest algorithm, with an accuracy rate of 99.55%.

In this paper, it is observed that higher accuracy rate can be achieved by using different features with examining previous studies. Different from previous studies, a new study based on features selected and coded from more features was made. 58 features were determined by making URL analysis. Using the machine learning method, the accuracy rates and model training times of different algorithms were compared.