IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS URL DETECTION BASED ON MACHINE LEARNING

[1]Mrs Dr. J. Sarada
[1]Associate Professor
[1]Department of Computer Applications
[1]Chadalawada Ramanamma Engineering College (Autonomous),Tirupathi

[2] K. Bhuvaneswari
[2] Student
[2] Department of Computer Applications
[2] Chadalawada Ramanamma Engineering College (Autonomous), Tirupathi

ABSTRACT

In recent years, the digital world has advanced significantly, particularly on the Internet, which is critical given that many of our activities are now conducted online. As a result of attackers' inventive techniques, the risk of a cyberattack is rising rapidly. One of the most critical attacks is the malicious URL intended to extract unsolicited information by mainly tricking inexperienced end users, resulting in compromising the user's system and causing losses of billions of dollars each year. As a result, securing websites is becoming more critical. In this paper, we provide an extensive literature review highlighting the main techniques used to detect malicious URLs that are based on machine learning models, taking into consideration the limitations in the literature, detection technologies, feature types, and the datasets used. Moreover, due to the lack of studies related to malicious Arabic website detection, we highlight the directions of studies in this context. Finally, as a result of the analysis, we conducted on the selected studies, we present challenges that might degrade the quality of malicious URL detectors, along with possible solutions.

Keywords:- URL, malicious URL detection, Feature extraction, feature selection ,machine learning

1.INTRODUCTION

Uniform Resource Locator (URL) is used to refer to resources on the Internet. In Sahoo et al. presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download. Malicious URLs can also be hidden

in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include Drive-by Download, Phishing and Social Engineering, and Spam.

According to statistics presented in, in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, botnet URLs, and phishing URLs, increase in number of attacks as well as danger level.

From the statistics of the increase in the number of malicious URL distributions over the consecutive years, it is clear that there is a need to study and apply techniques or methods to detect and prevent these malicious URLs.

Regarding the problem of detecting malicious URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques The method of detecting malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs. However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. The method of detecting malicious URLs based on behavior analysis techniques adopt machine learning or deep learning algorithms to classify URLs based on their behaviors. In this paper, machine learning algorithms are utilized to classify URLs based on their attributes. The paper also includes a new URL attribute extraction method.

In our research, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

The paper is organized as follows. Section II reviews some recent works in the literature on malicious URL detection. The proposed malicious URLs detection system using machine learning is presented in Section III. In this section, the new features for URLs detection process are also described in details. Experimental results and discussions are provided in Section IV. The paper is concluded by Section V.

2. LITERATURE SURVEY

In our research, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

3. PROBLEM STATEMENT

3.1 EXISTING SYSTEM

A. Signature based Malicious URL Detection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database query is executed. If the URL is blacklisted, it is considered as malicious, and then, a warning will be generated; otherwise URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list

B. **Machine Learning based Malicious URL Detection**

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semisupervised learning. And the detection methods are based on URL behaviors. In [1], a number of malicious URL systems based on machine learning algorithms have been investigated. Those machine learing algorithms include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, ect. In this paper, the two algorithms, RF and SVM, are used. The accuracy of these two algorithms with different parameters setups will be presented in the experimental results.

The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute groups are investigated, including Character and semantic groups; Abnormal group in websites and Host-based group; Correlated group.

3.2 DISADVANTAGE OF EXISTINTG SYSTEM:

- This method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules
- The attackers using spreading malicious URLs, Increases in number of attacks as well as danger level.

4.PROPOSED SYSTEM

4.1 PROPOSED SYSTEM:

In the proposed system, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

4.2 ADVANTAGE OF PROPOSED SYSTEM:

- This method is capable of detecting new malicious URLs that are in the set of predefined signs or rules .
- In the proposed work, SVM and RF are selected as an example to illustrate the good performance of the whole detection system, and are not our main focus.
- Readers are encouraged to implement some other algorithms such as Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc.

5.IMPLEMENTATION

5.1 Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse URLs Datasets and Train & Test Data Sets, View Urls Datasets Trained and Tested Accuracy in Bar Chart, View Urls Datasets Trained and Tested Accuracy Results, View Prediction Of Urls Type, View Urls Type Ratio, Download Predicted Data Sets, View Urls Type Ratio Results, View All Remote Users

5.2 View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users

5.3 Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN,PREDICT URLs TYPE,VIEW YOUR PROFILE.

6.CONCLUSION

In this paper, a method for malicious URL detection using machine learning is presented. The empirical results in Tables V and VI have shown the effectiveness of the proposed extracted attributes. In this study, we do not use special attributes, nor do we seek to create huge datasets to improve the accuracy of the system as many other traditional publications. Here, the combination between easy-to-calculate attributes and big data processing technologies to ensure the balance of the two factors is the processing time and accuracy of the system. The results of this research can be applied and implemented in information security technologies in information security systems. The results of this article have been used to build a free tool to detect malicious URLs on web browsers.

7. FUTURE ENHANCEMENT

In the proposed system, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature.

8.REFERENCES

D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey". CoRR, abs/1701.07179, 2017.

M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drivebydownload attacks and malicious javascript code," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281– 290.

R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24- 2019-en.pdf [Last accessed 10/2019].

S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian., 2008, pp. 91–96.

S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on., 2008, pp. 57–64.

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. ACM, 2009, pp. 681–688.

B. Eshete, A. Villafiorita, and K. Weldemariam, "Binspect: Holistic analysis and detection of malicious web pages," in Security and Privacy in Communication Networks. Springer, 2013, pp. 149–166.

S. Purkait, "Phishing counter measures and their effectiveness– literature review," Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

Y. Tao, "Suspicious url and device detection by log mining," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2014.

G. Canfora, E. Medvet, F. Mercaldo, and C. A. Visaggio, "Detection of malicious web pages using system calls sequences," in Availability, Reliability, and Security in Information Systems. Springer, 2014, pp. 226–238.

Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).

Thomas G. Dietterich. Ensemble Methods in Machine Learning. International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.

Developer Information. https://www.phishtank.com/developer_info.php. [Last accessed 11/2019].

URLhaus Database Dump. https://urlhaus.abuse.ch/downloads/csv/. [Ngày truy nhập 11/2019].

Dataset URL. http://downloads.majestic.com/majestic_million.csv. [Last accessed 10/2019].

Malicious_n_Non-MaliciousURL. https://www.kaggle.com/antonyj453/ urldataset#data.csv. [Last accessed 11/2019].

chrome.zip. ttps://drive.google.com/file/d/13G_Ndr4hMFx_qWyTEjHuOyJmHFW D0Gud/view?fbclid=IwAR0SLVCrvjHHGmoHZH97nXN3Bm- MY7jG4SOsKZYLAZjTFgeoJADfli64-g. [Last accessed 12/2019].