International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Detection of Malicious URLs using Classification Algorithm

Muskan V. Jaiswal[1], Dr. Anjali B. Raut[2]

[1]M.E. Scholar Department of Computer Science & Engineering H.V.P.M. C.O.E.T, Amravati, India
muskanjaiswal7697[at]gmail.com

[2]H.O.D Department of Computer Science & Engineering H.V.P.M. C.O.E.T, Amravati, India
anjali_dahake[at]rediffmail.com

Abstract: ***Malicious URL, a.k.a. malicious website, is a common and serious threat to cyber security. Malicious URLs host unsolicited content (spam, phishing, drive-by downloads, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner. Traditionally, this detection is done mostly through the usage of blacklists. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. This article aims to provide a comprehensive survey and a structural understanding of Malicious URL Detection techniques using machine learning. We present the formal formulation of Malicious URL Detection as a machine learning task, and categorize and review the contributions of literature studies that addresses different dimensions of this problem (feature representation, algorithm design, etc.). Further, this article provides a timely and comprehensive survey for a range of different audiences, not only for machine learning researchers and engineers in academia, but also for professionals and practitioners in cyber security industry, to help them understand the state of the art and facilitate their own research and practical applications. We also discuss practical issues in system design, open research challenges, and point out important directions for future research.***

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 [1], 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 [2, 3, 4]: Drive-by Download, Phishing and Social Engineering, and Spam.

According to statistics presented in [5], 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 [1, 2]. 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.

IIThe proposed method is outlined in Fig.1 and comprises three main components: pattern extraction, and alias extractionand ranking. Using a seed list of name-alias pairs, first there will be extraction of lexical patterns that are

Licensed Under Creative Commons Attribution CC BY

frequently used toconvey information related to aliases on the web. Theextracted patterns are then used to find candidate aliases fora given name. Various ranking scores can be defined using thehyperlink structure on the web and page counts retrievedfrom a search engine to identify the correct aliases amongthe extracted candidates.

2. Related Work

A. Signature based Malicious URL

Detection Studies on malicious URL detection using the signature sets had been investigated and applied long time ago [6, 7, 8]. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database queries isexecuted. 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 [9, 10, 11] 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 [12, 13]. 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

C. Malicious URL Detection Tools

- URL Void: URL Void is a URL checking program using multiple engines and blacklists of domains. Some examples of URL Void are Google SafeBrowsing, Norton SafeWeb and MyWOT. The advantage of the Void URL tool is its compatibility with many different browsers as well as it can support many other testing services. The main disadvantage of the Void URL tool is that the malicious URL detection process relies heavily on a given set of signatures.
- UnMask Parasites: Unmask Parasites is a URL testing tool by downloading provided links, parsing Hypertext Markup Language (HTML) codes, especially external links, iframes and JavaScript. The advantage of this tool is that it can detect iframe fast and accurately. However, this tool is only useful if the user has suspected something strange happening on their sites
- Dr. Web Anti-Virus Link Checker: Dr.Web Anti-Virus Link Checker is an add-on for Chrome, Firefox, Opera, and IE to automatically find and scan malicious content on a download link on all social networking links such as Facebook, Vk.com, Google+.
- Comodo Site Inspector: This is a malware and security hole detection tool. This helps users check URLs or enables webmasters to set up daily checks by downloading all the specified sites. and run them in a sandbox browser environment.
- Some other tools: Among aforementioned typical tools, there are some other URL checking tools, such as UnShorten.it, VirusTotal, Norton Safe Web, SiteAdvisor (by McAfee), Sucuri, Browser Defender, Online Link Scan, and Google Safe Browsing Diagnostic.

From the analysis and evaluation of malicious URL detection tools presented above, it is found that the majority of current malicious URL detection tools are signature-based URL detection systems. Therefore, the effectiveness of these tools is limited.

3. 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 [20] to detect malicious URLs on web browsers.

References

[1] Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.
[2] M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.
[3] M. Cova, C. Kruegel, and G. Vigna, “Detection and analysis of driveby- download attacks and malicious javascript code,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281– 290.
[4] 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.
[5] Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-242019-en.pdf [Last accessed 10/2019].

Licensed Under Creative Commons Attribution CC BY

[6] 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.
[7] 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. IEEE, 2008, pp. 57–64.
[8] 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.
[9] 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.
[10] S. Purkait, "Phishing counter measures and their effectiveness– literature review," Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

Licensed Under Creative Commons Attribution CC BY