International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 24-28

Malicious URL Detection using Logistic Regression

Shreyans Deshmukh
Computer Science and Engineering
Prof.Ram Meghe Institute of Technology & Research
Amravati,India

Vishal Rathod
Computer Science and Engineering
Prof.Ram Meghe Institute of Technology & Research
Amravati,India

Bhushan Sangani
Computer Science and Engineering
Prof.Ram Meghe Institute of Technology & Research
Darwha,India

Ms.R.A.Kale
Computer Science and Engineering
Prof.Ram Meghe Institute of Technology & Research
Amravati,India

Shubham Wankhade
Computer Science and Engineering
Prof.Ram Meghe Institute of Technology & Research
Wardha,India

***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, logistic regression, data, algorithm, malicious, benign, machine learning, analysis*

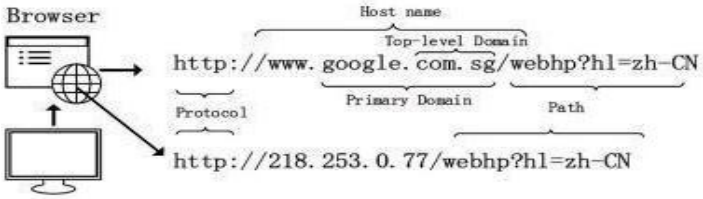
Date of Submission: 13-05-2022 Date of acceptance: 27-05-2022

I. Introduction

The advent of new communication technologies has had tremendous impact in the growth and promotion of businesses spanning across many applications including online- banking, e-commerce, and social networking. In fact, in today's age it is almost mandatory to have an online presence to run a successful venture. As a result, the importance of the World Wide Web has continuously been increasing.

Unfortunately, the technological advancements come coupled with new sophisticated techniques to attack and scam users. Such attacks include rogue websites that sell counterfeit goods, financial fraud by tricking users into revealing sensitive information which eventually lead to theft of money or identity, or even installing malware in the user's system. There are a wide variety of techniques to implement such attacks, such as explicit hacking attempts, drive- by download, social engineering, phishing, watering hole, man-in-the middle, SQL injections, loss/theft of devices, denial of service, distributed denial of service, and many others.

Considering the variety of attacks, potentially new attack types, and the innumerable contexts in which such attacks can appear, it is hard to design robust systems to detect cyber- security breaches. The limitations of traditional security management technologies are becoming more and more serious given this exponential growth of new security threats, rapid changes of new IT technologies, and



A. Fig 1.1 Components of URL

significant shortage of security professionals. Most of these attacking techniques are realized through spreading compromised URLs (or the spreading of such URLs forms a critical part of the attacking operation. URL is the abbreviation of Uniform Resource Locator, which is the global address of documents and other resources on the World Wide Web.

Compromised URLs that are used for cyber attacks are termed as malicious URLs. In fact, it was noted that close to one-third of all websites are potentially malicious in nature, demonstrating rampant use of malicious URLs to perpetrate cyber-crimes. A Malicious URL or a malicious web site hosts a variety of unsolicited content in the form of spam, phishing, or drive-by download in order to launch attacks. Unsuspecting users visit such web sites and become victims of various types of scams, including monetary loss, theft of private information (identity, credit-cards, etc.), and malware installation.

Phishing and Social Engineering, and Spam . Drive-by download refers to the (unintentional) download of malware upon just visiting a URL. Such attacks are usually carried out by exploiting vulnerabilities in plugins or inserting malicious code through JavaScript. Phishing and Social Engineering attacks trick the users into revealing private or sensitive information by pretending to be genuine web pages. Spam is the usage of unsolicited messages for the purpose of advertising or phishing. These attacks occur in large numbers and have caused billions of dollars' worth of damage, some even exploiting natural disasters. Effective systems to detect such malicious URLs in a timely manner can greatly help to counter large number of and a variety of cyber-security threats. Consequently, researchers and practitioners have worked to design effective solutions for Malicious URL Detection.

The most common method to detect malicious URLs deployed by many antivirus groups is the blacklist method. Blacklists are essentially a database of URLs that have been confirmed to be malicious in the past. This database is compiled over time (often through crowd-sourcing solutions, e.g. Phish Tank), as and when it becomes known that a URL is malicious. Such a technique is extremely fast due to a simple query overhead, and hence is very easy to implement. Additionally, such a technique would (intuitively) have a very low false-positive rate (although, it was reported that often blacklisting suffered from non-trivial false-positive rates. However, it is almost impossible to maintain an exhaustive list of malicious URLs, especially since new URLs are generated every day. Popular types of attacks using malicious URLs include: Drive-by Download, Attackers use creative techniques to evade blacklists and fool users by modifying the URL to ―appear" legitimate via obfuscation. B. Eshete, A. Villafiorita, and K. Weldemariam et. al. identified four types of obfuscation: Obfuscating the Host with an IP, Obfuscating the Host with another domain, Ob fuscating the host with large host names, and misspelling. All of these try to hide the malicious intentions of the website by masking the malicious URL.

Recently, with the increasing popularity of URL shortening services, it has become a new and widespread obfuscation technique (hiding the malicious URL behind a short URL). Once the URLs appear legitimate, users visit them, and an attack can be launched. This is often done by malicious code embedded into the JavaScript. Often attackers will try to obfuscate the code so as to prevent signature based tools from detecting them. Attackers use many other techniques to evade blacklists including: fast-flux, in which proxies are automatically generated to host the webpage; algorithmic generation of new URLs; etc. Additionally, attackers can simultaneously launch more than one attack to alter the attack-signature, making it undetectable by tools that focus on specific signatures. Blacklisting methods, thus have severe limitations, and it appears almost trivial to bypass them, especially because blacklists are useless for making predictions on new URLs.

To overcome these issues, in the last decade, researchers have applied machine learning techniques for Malicious URL Detection Machine Learning approaches, use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious or benign. This gives them the ability to generalize to new URLs unlike blacklisting methods. The primary requirement for training a machine learning model is the presence of training data. In the context of malicious URL detection, this would correspond to a set of large number of URLs. Machine learning can broadly be classified into supervised, unsupervised, and semi-supervised, which correspond to having the labels for the training data, not having the labels, and having labels for limited fraction of training data, respectively. Labels correspond to the knowledge that a URL is malicious or benign.

II. LITERATURE REVIEW

Many researchers have proposed different methods for classification and detection of malicious Web pages and detection of different Webpage attacks. D. Sahoo, C. Liu et al, In 2017 have described how a machine can able to judge the URLs based upon the given feature set. Specifically, they described the feature sets and an approach for classifying the given the feature set for malicious URL detection as the traditional methods described above falls short in detecting the new malicious URLs on its own[1].

To counter these limitations, J. Ma, L. K. Saul, S. Savage, and G. M. Voelker et al proposed a novel approach using sophisticated machine learning techniques that could be used as a common platform by the Internet users in order to detect the malicious URLs. Various feature sets for the URL detection have also been proposed that can be used with Support Vector Machines (SVM). The feature set used in is composed of the 18 features, such as token count, average path token, largest path, largest token, etc. They also propose a generic framework that can be used at the network edge. That would safeguard the naive users of the network against the cyber-attacks.

Although, using this method of detecting malicious URLs based on various features did not give much accuracy and obtaining features with a high collection time maybe infeasible.

The comparison has been made on the various machine learning techniques. The detailed view of the results of various machine learning techniques has been elaborated in. Machine Learning approaches use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious. This gives them the ability to generalize to new URLs unlike blacklisting methods.

In, Internet Security Threat Report (ISTR) 2019–Symantec et al conducted a comprehensive and systematic survey on Malicious URL Detection using machine learning techniques. In this survey, they categorized most, if not all, the existing contributions for malicious URL detection in literature, and also identified the requirements and challenges for developing Malicious URL detection as a service for real-world cyber-security applications.

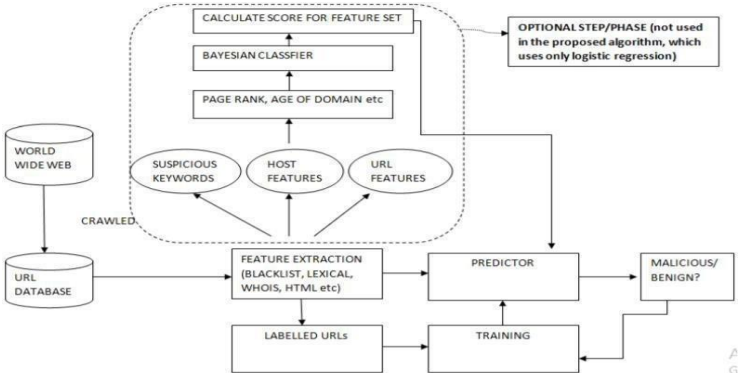
M. Cova, C. Kruegel, and G. Vigna eternal performed an extensive literature survey of existing techniques and approaches for malicious Web pages detection. Presents a brief overview of various forms of Web pages attacks. M. Cova, C. Kruegel, and G. Vigna et al introduced different Web pages and URLs features used for the effective detection of the malicious Web pages and also online learning algorithms as a promising approach for the large scale and efficient detection of malicious Web pages.

Keywords, host features, URL features by using extraction techniques like Blacklist, Lexical, WHOIS, HTML, etc. The host features consists of Page Rank, Age of Domain, etc. and then score is calculated for the feature set using Bayesian.

In considering limitations of previous work for malicious URLs detection based on key features like URL features, URL source features, domain name features and short URLs features, M. Cova, C. Kruegel, and G. Vigna et al proposed a methodology to detect malicious URLs and identify attack types. 117 various types of discriminative features like URL features, domain name features, URL source features and short URLs features were used. Significant results were obtained by using proposed novel features.

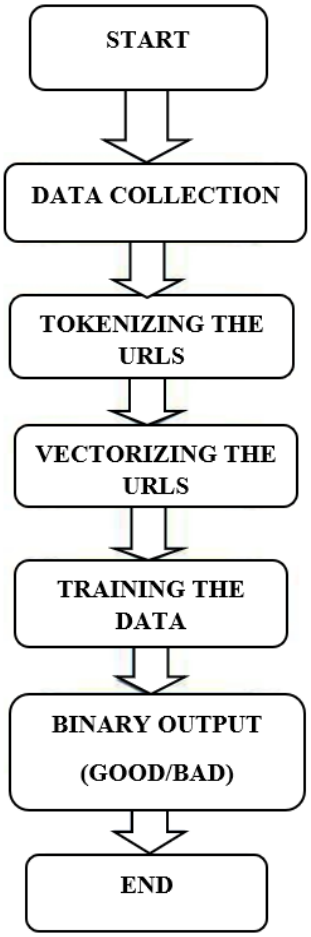
Although, there was still need to investigate more features of short URLs for the effective detection and attack type identification, because it is the most growing trend today on the micro blogging sites like Twitter, Facebook etc.

The first task is gathering data. Some websites offer malicious links while browsing. The next task is finding out clear URLs. We can use datasets which are already available so there is no need to crawl for non-malicious URLs. The feature extraction is used which extracts suspicious Classifier. Then the links are checked that if they are malicious or not.



B. Fig. 2.1 Malicious URL Detection

FLOWCHART OF PROPOSED MODEL--



REFERENCES

[1]. M. Khonji, Y. Iraqi, and A.Jones,―Phishing detection: a literature survey,‖ IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.
[2]. 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.
[3]. R. Heartfield and G.Loukas―Ataxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks,‖ ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.
[4]. Internet Security Threat Report (ISTR) 2019–Symantec.
[5]. https://www.symantec.com/content/dam/symantec/docs/reports/istr-242019-en.pdf [Last accessed 10/2019].
[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.
[8]. MALWARE 2008. 3rd International Conference on.IEEE, 2008, pp. 57–64
[9]. 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