IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identification and Prevention approaches for Web-based Attacks using Machine Learning Techniques

[1]Aasha Singh, [2]Dr. Awadhesh Kumar, [3]Dr. Ajay Kumar Bharti

[1]Research Scholar, [2]Associate Professor, [3]Associate Professor

[1]Computer Science & Engineering

[1]MUIT, Lucknow, India

Abstract: Identification and prevention from web-based attacks is the need of any organization which deals with the important information about their employee and the work done by the employees. The information is the backbone of any organization and it may be a disaster when the information could be stolen. There are several factors which may affect the identification of malicious Web Contents. First one is the large quantity web data or scale of the web that can create problem. Second is the nature of data on the web. Since this heterogeneous and complex nature of data may become the problem in identification of benign websites and malicious websites. Another aspect of attacking on the multiple websites by changing their domain names, it may be very easy for attackers to attack on multiple and unexpected locations. There are many solutions provided by researchers which are providing security from various types of attacks using malicious Web Contents.

***Index Terms* – WWW, URLs, Web Content, Attack, Virus, Phishing.**

I. INTRODUCTION

As we can see that World Wide Web is being growing day by day and Internet have converted into the basic need of each and every one. As the growth of Internet increasing, the malicious software like malware or attacks are transmitted. The limitations of traditional security measures and management are creating problems in identification and prevention of various types of attacks. Attackers may design different types of malicious Web Contents to get the information from the particular organization or from the particular person who is having precious information. Active systems which can discover such malevolent URLs routinely can seriously support to hostage huge amount of and a diversity of cyber-threats, intimidations.

1.1 URL Detection

URL detection searches and matches the entire Internet traffic belongs to a well-defined database, and then allows or rejects access to a website based on the database either it exists in the database or not.. **A URL identification** database will allocate classified websites to a URL category, or a group. This will also describe the situations of entrée to that URL. As given example, an address might be [17]:

- **Blocked:** it is defined on a website by website base. This can put on to disturbing sites, like local news or social media or websites that are known to hold different kinds of malware.
- **Allowed:** they are SaaS sites. They are pertinent to an organization and workflow of that organization.
- **Involved to defined IT policies:** they could seek to a specific site and could be listed and planned, that is why IT tools can check that who visits particular websites and the visiting time also.
- **Blocked and allowed URLs categories:** when the activities are performed not on a website-by-website base, but somewhat by the type encircling multiple websites. This might have classes of malware or phishing websites, distracting websites, and questionable websites.

The databases of URL detection filters can be warehoused on the places or live in the cloud, or might be both, it depends on the system's need. Native lookups could help to decrease latency between the filter and user, if websites are regularly visited. On the other hand, a cloud database can be trusted upon to sustain an latest log of all identified websites. In a fusion result, different devices can acclimate to exceptional traffic forms, and practice known users' traffic to accumulate more newly opened URLs in the on-device cache memory, decreasing latency. As per the need, a master database which is placed in the cloud can be inquired when the websites are not found in the device local cache. Preferably, websites are categorized automatically.

1.2 Web Content

Web Content known for to the written, audio, or graphic content available on a website. Content means some ingenious component, for illustration, script, submissions, pictures, stored e-mail messages, e-services, data, video and audio files. Detection of Web Content is the procedure of a application program to monitor and/or omit entrée to a web page or an email supposed to be offensive. Detection of Web Content is used by organizations like they use their firewalls and by personal users. It helps as identifying Web Content configurations like objects or text strings within an image, if relates, specify objectionable Web Content that is to be separated out. Then filter of Malicious Web Content will stop the access to that Web Content. Offensive, unsuitable, or unlawful Web Content makes risk for officialdoms. Web Content identification and detection benefits to lessen these threats by creating such Web Content hard to contact in the office, and by proving the company's prejudice for unsuitable, unlawful, or offensive Web Content in general [17].

1.3 Difference between URL Detection and Web Content Detection

URL blocking and Web Content detection offer an additional whole Internet access control solution to the user. The mainstream of unwanted traffic is omitted rapidly and professionally by URL blocking on the outbound pathway. The rest is trapped on the coming back path. The worker also has the suppleness to permit entrée to a website but selectively chunk certain Web Content from that website. Phishing, Drive-by Download, Spam, and Social Business are utmost prevalent kinds of attacks by means of malicious URLs. By misusing susceptibilities in plugins or implanting malicious code over JavaScript, Drive-by-download attack is usually carried out. For upsetting the sincere webpages, the social engineering and phishing attacks idea users into revealing their delicate material. Spam is cast-off as unpaid messages that are used for publicizing or phishing. Each year these varieties of attacks lead numerous difficulties. So the foremost fear exists today is discovering such malicious URLs and Web Content in a well-timed manner. New features of websites data (such as huge scale, great measurement, sparsity, and varying forms) remember interesting the traditional detecting approaches [17].

II. TYPES OF CYBER-ATTACKS

These are the most common cyber-attacks types are given below:

- **Injection attacks**: the SQL injection method is the most popular way of attacking by attackers. This method targets the websites and directly to the database of the server.
- **Spamming:**
- **Phishing:** this is the most common method of attacks done by cyber attackers. This method is easy to perform and effective too.
- **Spoofing:** when the attackers satirize another device or a person on the network to target the attacks against the network hosts to steal information using bypassing access controls.
- **Denial-of-service (DoS):** this type of attack controls the system's resources so that system could not reply to the service requests. A dispersed denial-of service is a kind of violence on the resources and systems, but then again this attack is thrown from various number of different user machines that are infested by malevolent program under control of attackers.
- **Drive–by download attack:** these types of attacks are very general method of distribution malware. Attackers search for doubtful websites and put a malevolent script in the PHP or HTTP code on one of the pages. It might be happen when go to see a website or looking to a pop-up window or an e-mail.
- **Password attack:** this is the very easiest method to hack the information from the system of a person. Looking around the person may be one of the ways to obtain the password. "Sniffing" the network connection to obtain the unencrypted passwords using different ways like social engineering, to get access the password database or it might be guessing.
- **Brute-force:** in this way of guessing password means via a random way of try to use dissimilar passwords and some patterns can also be tried to crack the password.
- **Eavesdropping attack:** it might be possible when attackers intercept the web traffic. By using this an invader can get passwords, OTP's, credit cards data and other important information over the network.
- **Birthday attack:** this type of attacks is prepared contrary to hash algorithms. Hash algorithms can be used to verify the reliability of a message, digital signature or software. This attack is the probability of conclusion two random messages that produce the same message digest when processed by a hash function.
- **Malware attacks:** if undesirable program is set up in any system deprived of of consent or information, that software is called malicious software. This kind of software could assign itself to authentic code and spread; it an also prowl in advantageous applications or reproduce themself through the network.

The list of such common malwares are:

- **Micro virus:** A micro virus infects the application packages like MS Word and Excel. They are attached with the initialization sequence of an application. As the application starts, the virus enables the instructions prior to transfer control to that application. This virus reproduces itself and allocates to new code in the computer system.
- **File infectors:** these viruses habitually ascribe themselves to an executable code files like .exe files. As the code is loaded the virus becomes installed. A different version of a file infector associates itself with that file which is created by bug file with the similar name.
- **Polymorphic viruses:** this kind of viruses hides themselves through changing cycles of encryption and decryption. This virus harms to an area of code. The hidden virus with related mutation engine primarily is decrypted by a decryption program. It is very tough to detect such kind of viruses.
- **System or boot-record infectors:** a system or boot-record infector assigns to the master boot record on the hard disks. At the time of system boot it seeks the boot sector and propagates the virus into the memory and spreads to other disks and systems.
- **Trojans or Trojan horse:** this program embedded itself in a beneficial program having a malicious function. They do not replicate itself as the viruses do.
- **Stealth viruses:** they conquest system functions to hide themselves. The virus can do this by negotiating malware encounter software so that anti-virus software will claim an infected region as to be uninfected region.
- **Worms:** these are different and vary from viruses since they do not close to the host file. They are autonomous programs that disseminate across Internet and computers.
- **Logic bombs:** this is a kind of malicious software which is embedded to request and is initiated by a particular event like specific date or a logical condition.
- **Droppers:** these are the programs that are used to set up viruses on the computers. Since it might not have some malicious code so it might be possible that it could not be detected by antivirus software. It can also download and install updates for viruses reside in the compromised system.
- **Ransomware:** this is a malware of different type that stops gathering to the user's data and lurks to delete or publish it unless the deal is paid.
- **Spyware:** it is a kind of program used for collecting data about various users, their surfing habits and system information. This software tracks everything what a user do without his information and it transfers data to the distant user. One more special thing is that it can easily download and install malicious software.
- **Adware:** it is a special kind of software which companies use it for advertisements and marketing purposes.

III. DETECTION OF MALICIOUS URLS

Malicious Web Contents or URLs are the basic path for running of different cyber-attacks like spamming, phishing type attacks or malware attacks. Different malicious URLs are that URLs which are compromised URLs and used for cyber-attacks. Detection of malicious Web Contents and URLs is as important as identification of their attack types since identifying the threat enables us to estimate the severity of the loss done by attack and helps us to adopt an effective and preventive measure. The purpose of detection of malicious URLs is that an employee could avoid or ignore that malicious URLs (objectionable websites, illegal Web Contents, websites that are not related to work or websites associated with phishing attempts) when they are identified so that information could not be lost from the any personal computer or from an organization.

Various methodologies have been proposed to identify malicious URLs and Web Content. These are classified into four categories: Blacklists, Heuristics, URLs-based grouping, Web Content-based ordering, and different types of feature Selection based classification approaches [16]. To overcome the problems of Blacklists and Heuristics methods, various researchers had used machine learning algorithms to identify malicious URLs and Web Content from benign one.

It is known that Machine Learning algorithms and its techniques offer a system (1) the process of learning by itself (2) improving the result from past experience (3) No need of any definite program. The key objective of machine learning approach is to offer capability to the computer system to acquire knowledge automatically deprived of any interfering with individuals.

Machine learning approaches perform the following tasks [17]:

1. To construct a model, the numerous algorithm of machine learning is trained by means of a set of particular amount of training data.
2. As the fresh input data is go into the algorithm of machine learning, this it proceeds certain prediction on the base of the trained model.
3. The process of prediction carried out in stage 2 could be calculated for inspection accuracy.
4. When the assessed accurateness is just the once tolerable, then a particular algorithm of machine learning can be applied. In case of not happening this an improved set of training data is used for training of machine learning algorithm again and again.

There is various URL detection methods are proposed by different researchers. Some researchers had proposed method for detection of malicious URLs of a single type of attack using machine learning methods and some had proposed for almost all popular attack types using machine learning methods.

IV. DETECTION OF MALICIOUS URLS FOR SINGLE ATTACK

Machine learning approach can be used for detection of malicious Web Contents and URLs. These are some approaches that had been proposed by various researchers for single attack type.

- In [3], authors proposed a malicious Web Content detection method applying machine learning.
- In [2], authors proposed a statistical approach to classify phishing emails.
- In [4], authors proposed a spam webpages detection scheme using content analysis. They had applied a website depend heuristics approach, like words written in a particular page or label and also the portion of observable content.
- In [5], authors proposed an analysis of malignity of a huge group of webpages applying machine learning approach for VM-based analysis like a pre-filter.
- In [6], here authors proposed a classifier of phishing website for updating the Google's phishing blacklist inevitably. They utilized various features gained from province statistics and webpage contents.
- In [7], proposed a framework for spam signature generation known as AutoRE for detection of spam emails using botnet approach. This AutoRE takes URLs as input in emails and produce regular expression signatures as outputs which could detect botnet spam.

V. DETECTION OF MALICIOUS URLS WITH MULTIPLE ATTACK TYPES

Mostly existing approaches which are based on machine learning habitually focus on a on its own type of attack and malevolent behaviour. These approaches generally use machine learning technique to adjust their classification models. There are some approaches that had been proposed by various researchers for multiple attack types.

- The authors in [8, 9], proposed a classification model that can identify spam URLs and various phishing URLs. They have defined a classification technique for URLs applying statistical methods on the lexical as well as host-based characteristics of malicious URLs.
- Many researchers have proposed a web spams detection schemes by applying propagating trust or disbelief via links [10], detection of bursts of linking action as a doubtful signal [11].
- In [13], authors proposed web spams detection schemes using integrating link and content-based features.

REFERENCES

[1] Hyunsang C., Bin B. Zhu, And Heejo L. 2011. Detecting Malicious Web Links and Identifying Their Attack Types. Usenex Conference on Web Application Development.

[2] Fette, I., Sadeh, N., And Tomasic A. 2007. Learning to detect phishing emails. In WWW: Proceedings of the international conference on World Wide Web.

[3] Hou, Y.-T., Chang, Y., Chen, T., Laih, C.-S. 2010. AND CHEN, C.-M. Malicious Web Content detection by machine learning. Expert Systems with Applications, 55–60.

[4] Ntoulas, A., Najork, M., Manasse, M., And Fetterly, D. 2006. Detecting spam web pages through content analysis. In WWW: Proceedings of international conference on World Wide Web.

[5] Provos, N., Mavrommatis, P., Rajab, M. A., And Monrose.,F. 2008. All your iFRAMEs point to us. In Security: Proceedings of the USENIX Security Symposium.

[6] Whittaker, C., Ryner, B., And Nazif, M. 2010. Large-scale automatic classification of phishing pages. In NDSS: Proceedings of the Symposium on Network and Distributed System Security.

[7] Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., And Osipkov, I. 2008. Spamming botnets: signatures and characteristics. In SIGCOMM.

[8] Ma, J., Saul, L. K., Savage, S., And Voelker, G. M. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In KDD: Proceedings of the international conference on Knowledge Discovery and Data mining.

[9] Ma, J., Saul, L. K., Savage, S., And Voelker, G. M. 2009. Identifying suspicious URLs: an application of large-scale online learning. In ICML: Proceedings of the International Conference on Machine Learning.

[10] Gy¨Ongyi, Z., And Garcia-Molina, H. 2005. Link spam alliances. In VLDB: Proceedings of the international conference on Very Large Data Bases.

[11] Shen, G., Gao, B., Liu, T.-Y., Feng, G., Song, S. 2006. AND LI, H. Detecting link spam using temporal information. IEEE International Conference on Data Mining, 1049–1053.

[**12**] Castillo, C., Donato, D., Gionis, A., Murdock, V., And Silvestri, F. 2007. Know your neighbors: web spam detection using the web topology. In ACM SIGIR: Proceedings of the conference on Research and development in Information Retrieval.

[13] Chung, Y.-J., Toyoda, M., And Kitsuregawa, M. 2010. Identifying spam link generators for monitoring emerging web spam. In WICOW: Proceedings of the 4th workshop on Information credibility.

[**14**] https://networkinterview.com/url-filtering-vs-content-filtering/

[15] Tie Li A, Gang Kou B, Yi Peng A. 2020. Improving Malicious Urls Detection Via Feature Engineering: Linear And Nonlinear Space Transformation Methods. Information Systems 91 (2020) 101494, Https://Doi.Org/10.1016/J.Is.2020.101494, Published By Elsevier Ltd.

[16] Justin Ma, Lawrence K Saul, Stefan S., and Geoffrey M.V. 2009. Identifying suspicious URLs: an application of large-scale online learning. In Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 681–688.