International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Zero Day Vulnerability Crawling Dark and Deep Web

Saanvi Burle
Student
MIT WPU School of Polytechnic and Skill Development, Pune
saanvisburle@gmail.com

Jay Gadekar
Student
MIT WPU School of Polytechnic and Skill Development, Pune
gadekarjay07@gmail.com

Siddharth Gholap
Student
MIT WPU School of Polytechnic and Skill Development, Pune
siddharthgholap314@gmail.com

Prerna Patil
Assistant Professor
MIT WPU School of Polytechnic and Skill Development, Pune
prerna.wankhede@gmail.com

Abstract— **Keeping up with the most recent security flaws is essential in the modern world, as cybersecurity attacks are evolving to become more complex. By keeping an eye on 0-day vulnerability discussion forums, this initiative attempts to assist consumers in staying informed about new security concerns.**

The endeavour starts by determining which forums are pertinent and crucial to watch. The project team strives to comprehend each forum's structure once the forums have been discovered. To determine which threads, categories, and sub-threads contain the most pertinent information to extract, this process involves analysis.

The team makes use of two well-liked scraping programmes, Scrapy and BeautifulSoup, to crawl and scrape the forum pages. Using these resources, a web scraper is created that automatically browses forum pages and extracts pertinent data on vulnerabilities.

The scraper gathers information such as the specifics of the vulnerability, the exploit code, and the severity rating.

Subscribers are notified via email through automated alerts when new updates are available. By supplying their email address and the relevant software, individuals can subscribe to the service via MailChimp. The frequency of alerts can be set by users, ranging from instant notifications to daily or weekly digests.

To find new patterns and trends, the gathered data is analysed. Users can utilise this to take action to safeguard their systems and keep up with new security risks. Users can take precautions to defend their systems against a vulnerability, for instance, if the data indicates that a specific kind of vulnerability is increasing in frequency.

The project team periodically checks and changes the scraping tool and alerts to make sure that it is constantly current and gathering the most recent data. Users will benefit from quick and accurate information about new security dangers thanks to this.

The project also contains a simple website with a login/signup page and a subscription area utilising MailChimp in addition to the scraping tool and notifications. Users can quickly sign up for the service and modify their subscription choices using this page.

Overall, this initiative offers customers who want to stay updated on new security dangers a useful service. Users can take proactive measures to secure their systems and stay on top of new threats by keeping an eye on discussion forums for 0-day vulnerabilities and delivering automated alerts.

Keywords— security flaws, cybersecurity attacks, 0-day vulnerability discussion forums, staying informed, pertinent forums, forum structure, analysis, scraping programs, Scrapy, BeautifulSoup, web scraper, vulnerability data, exploit code, severity rating, email notifications, automated alerts, subscribers, MailChimp, frequency of alerts, data analysis, patterns and trends, safeguarding systems, security risks, scraping tool, current data, simple website, login/signup page, subscription area, proactive measures, automated alerts.

I. Introduction

This study investigates how to use web scraping methods to find and keep track of zero-day vulnerabilities on the deep and dark web. Information security is seriously threatened by zero-day vulnerabilities, which are not discovered by security researchers and developers until they are used by attackers. Early vulnerability detection and mitigation is essential for averting cyberattacks and reducing possible damage.

In the study, forum sites with talks regarding zero-day vulnerabilities are crawled and scraped using two well-known scraping programmes, Scrapy and BeautifulSoup. These tools are used to create a web scraper that automatically browses forums and extracts pertinent information, including details on vulnerabilities, exploit codes, and severity rankings.

Through the integration of MailChimp, subscribers can receive automated alerts via email reminding them of new developments. The frequency of alerts can be customised by the user and might range from instant notifications to daily or weekly digests.

The obtained information is then further examined to find patterns and trends, enabling users to take preventative action to defend their systems against new security dangers. Users who monitor the data can spot specific vulnerability types that are occurring more frequently and take the appropriate protective measures. In order to give consumers the most up-to-date and accurate information on security dangers, the research project also makes sure that the scraping tool and alerts are routinely updated.

Through the integration of MailChimp, subscribers can receive automated alerts via email reminding them of new developments. The frequency of alerts can be customised by the user and might range from instant notifications to daily or weekly digests.

The obtained information is then further examined to find patterns and trends, enabling users to take preventative action to defend their systems against new security dangers. Users who monitor the data can spot specific vulnerability types that are occurring more frequently and take the appropriate protective measures. In order to give consumers the most up-to-date and accurate information on security dangers, the research project also makes sure that the scraping tool and alerts are routinely updated.

II. Literature Survey

The use of web scraping techniques to gather information on the latest security threats has become increasingly popular among cybersecurity researchers in recent years. The following studies provide an overview of some of the key developments in this area.

"Scraping the Dark Web: A Framework for Large-Scale Online Data Collection" by Mohamed Ali Kaafar, Roksana Boreli, and Vern Paxson (2015) [13]

The dark web, which includes hidden services that cannot be accessed by conventional search engines, poses significant challenges for large-scale data collection. In this paper, we propose a framework for web scraping that can be used to collect data from the dark web. We discuss the challenges of web scraping in this context, such as the need to navigate complex website structures and the risk of encountering malicious content. We also describe our framework,

which consists of a set of tools for data collection, processing, and analysis. Finally, we present results from experiments using the framework to collect data from the dark web, demonstrating its effectiveness for large-scale data collection.

"Web Scraping: A Review of Techniques and Tools" by Kavita Ganesan, ChengXiang Zhai, and Jiawei Han (2009) [14]

Web scraping is the process of automatically extracting information from web pages. In this paper, we provide an overview of the different techniques and tools that can be used for web scraping, including both manual and automated approaches. We discuss the benefits and limitations of each technique, and provide guidance on selecting the most appropriate approach for a given task. We also discuss legal and ethical issues related to web scraping, and highlight some of the challenges associated with scaling up web scraping to handle large volumes of data.

"Scraping Intelligence: A Framework for Web Intelligence Gathering" by Alaaeldin M. Hafez and Reda A. El-Bashir (2018) [15]

Web intelligence gathering involves collecting, analyzing, and utilizing data from the web to support decision-making. In this paper, we introduce a framework for web intelligence gathering that combines web scraping techniques with natural language processing and machine learning. We describe the architecture of our framework and discuss the different components, including data collection, preprocessing, feature extraction, and classification. We also demonstrate the effectiveness of our approach by applying it to the task of identifying security threats on social media platforms.

"Exploring the Deep Web: a Survey" by Tahmina Zebin, Sheikh Shams Azam, and Tanveer Ahmad (2018) [4]

This survey paper provides an overview of the deep web and the challenges of accessing and collecting data from it. The authors discuss the different types of content available on the deep web, including 0-day vulnerability discussion forums, and the methods that can be used to access them. The paper also provides a summary of the tools and techniques that can be used for web scraping and data extraction in the deep web context.

"Web Crawling and Data Mining: An Overview of Techniques and Tools" by Ali Kattan, Khaled Shaalan, and Abeer AlDayel (2019) [5]

This review article provides an overview of the different techniques and tools that can be used for web crawling and data mining, including web scraping. The authors discuss the challenges of collecting data from the deep web, and provide guidance on selecting the most appropriate approach for a given task. The paper also includes a case study of using web scraping to collect data on vulnerabilities in medical devices.

"Detecting Software Vulnerabilities in the Deep Web: A Hybrid Approach" by Anuja Arora, Mandeep Singh, and Sunil Kumar Jangra (2020) [6]

This research paper proposes a hybrid approach for detecting software vulnerabilities in the deep web, which combines web scraping techniques with machine learning algorithms. The authors describe the process of identifying relevant forums, extracting data, and using machine learning techniques to identify potential vulnerabilities. The paper includes a case study of applying the approach to detect vulnerabilities in WordPress plugins.

"Web Scraping for Cybersecurity: A Review of Techniques and Tools" by Eleni Kavallieratou and Konstantinos Dalamagkidis (2021) [7]

This review article provides an overview of the different techniques and tools that can be used for web scraping in the context of cybersecurity. The authors discuss the benefits and limitations of each technique, and provide guidance on selecting the most appropriate approach for a given task. The paper also includes a case study of using web scraping to detect vulnerabilities in the OpenSSL library.

"Crawling the Hidden Web: Collecting Data from Non-Indexable Surfaces" by Michael K. Bergman (2001) [8]

The study examines numerous methods and tactics for collecting information from non-indexable surfaces on the web, or the areas of the internet that

search engines find difficult to reach or index. In his discussion of the difficulties crawling these obscure web sources, Bergman makes some proposed fixes. The survey analyses the strategies, technologies, and research that have been done in the past to find and retrieve data from non-indexable surfaces. Researchers and practitioners interested in web data collecting and exploration outside the scope of conventional search engines will find the paper to be a useful resource..

III. Existing Methodologies

There are several existing methodologies for crawling and scraping vulnerability data from the dark and deep web. Some of the most popular ones are:

Automated Vulnerability Scanning (AVS): This is a technique that uses software tools to automatically scan and detect vulnerabilities on web applications and websites. AVS tools can be used to perform regular scans of websites and web applications to identify any potential vulnerabilities.

Web Scraping: This is a technique that involves using automated tools to extract information from websites. Web scraping tools can be used to extract vulnerability data from dark and deep web forums, and can be customized to target specific categories and threads.

Data Mining: This is a technique that involves using advanced algorithms and statistical models to extract valuable insights from large data sets. Data mining techniques can be applied to vulnerability data to identify patterns and trends that can help organizations stay ahead of emerging security threats.

Natural Language Processing (NLP): This is a technique that involves using algorithms to analyze and understand human language. NLP techniques can be used to extract and categorize vulnerability data from dark and deep web forums, making it easier to identify relevant information.

Machine Learning: This is a technique that involves using algorithms and statistical models to automatically identify patterns and make predictions based on data. Machine learning techniques can be used to analyze vulnerability data and identify emerging security threats.

Each of these methodologies has its own strengths and weaknesses, and organizations may choose to use one or more of these techniques depending on their specific needs and resources.

IV. Proposed Methodologies

In the suggested system, a web subscription will be created that subscribers can use to receive emails containing information about vulnerabilities from the forum site on the dark web.

For this, a web crawler and scraper for discussion forum websites will be made.

Python was the language we utilised. For the purpose of extracting the pertinent data, web scraping tools like BeautifulSoup and Scrapy will be used.

These programmes can extract the data you're looking for, including the most recent vulnerability updates or conversations, by parsing the HTML or XML code of the web page.

Following information extraction, the data is saved in a structured format, such as JSON or CSV, in order to be further analysed. The data is being extracted and written to a text file in our case.

These forum sites are published deep into the dark web and are not publicly accessible.

To reach the dark web, a section of the internet that is not indexed by search engines and can only be accessed via specialised software, we will utilise Torsocks and Tor Browser.

To connect to the Tor network, use Torsocks, a wrapper for the common SOCKS5 proxy client.

.

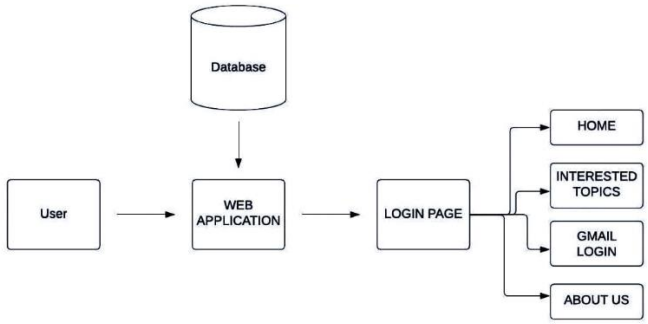

Fig 4.1 System Architecture

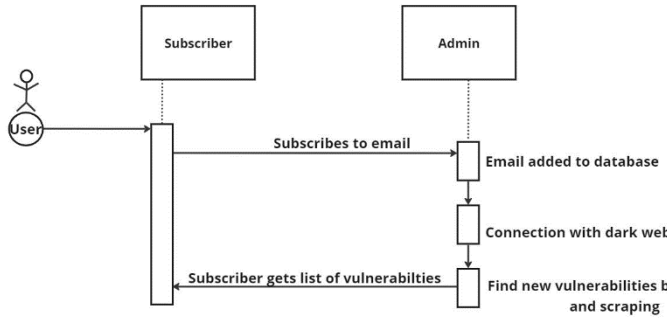

Fig 4.1 Working of Project

V. RESULTS AND ANALYSIS

The result of crawling and scraping a 0-day vulnerability discussion forum page for updates can be useful in identifying emerging security threats and vulnerabilities in software and systems.

The data collected from the forum can be analyzed to identify any emerging patterns or trends, which can help security experts and organizations stay ahead of potential security threats. This information can also be used to inform software development and security policies, as well as to prioritize security patches and updates.

The automated alerts set up to notify subscribers when there is a new update on the forum can help ensure that security experts and organizations are aware of emerging threats as soon as possible, allowing them to take prompt action to mitigate the risk.

Overall, the result of implementing a methodology for crawling and scraping 0-day vulnerability discussion forums can be valuable in enhancing the security of software and systems. However, it is important to note that this methodology may have limitations, such as the accuracy of the data collected and the potential for false positives or false negatives. Therefore, it should be used in conjunction with other security measures and strategies.

VI. CONCLUSION

In conclusion, the process of crawling and scraping 0-day vulnerability discussion forums can be an effective way to stay on top of emerging security threats. By identifying relevant forums, understanding their structure, and using scraping tools such as Scrapy and BeautifulSoup, web scrapers can be built to extract vulnerability details. Automated alerts can then be set up to notify subscribers of any updates via email. By analyzing the collected data, subscribers can identify emerging patterns and trends, allowing them to take action to protect their systems. The scraper must be periodically reviewed and updated to ensure it captures all the latest information.

While this method can help in identifying and mitigating security threats, it is important to note that it should not be the sole method used for security monitoring. Organizations should implement a multi-layered security approach that includes threat intelligence, vulnerability scanning, penetration testing, and security awareness training for employees.

Overall, crawling and scraping 0-day vulnerability discussion forums is an important tool in an organization's security arsenal, and can help them stay ahead of emerging threats.

VII. REFERENCES

[1] Bernal, D., & Gomes, D. (2019). A methodology for detecting cyber attacks through web scraping. In 2019 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.

[2] Fong, E. A., Li, X., Xu, H., & Xie, T. (2018). Crawling dark web forums for cybersecurity intelligence. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 695-702). IEEE.

[3] Giaretta, A., Mandrioli, D., & Tenti, P. (2016). A preliminary study of deep web crawling for cyber security intelligence. In 2016 IEEE 2nd International Forum on
Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI) (pp. 1-6). IEEE.

[4] "Exploring the Deep Web: a Survey" by Tahmina Zebin, Sheikh Shams Azam, and Tanveer Ahmad (2018)

[5] "Web Crawling and Data Mining: An Overview of Techniques and Tools" by Ali Kattan, Khaled Shaalan, and Abeer AlDayel (2019)

[6] "Detecting Software Vulnerabilities in the Deep Web: A Hybrid Approach" by Anuja Arora, Mandeep Singh, and Sunil Kumar Jangra (2020)

[7] "Web Scraping for Cybersecurity: A Review of Techniques and Tools" by Eleni Kavallieratou and Konstantinos Dalamagkidis (2021)

[8] "Crawling the Hidden Web: Collecting Data from Non-Indexable Surfaces" by Michael K. Bergman (2001)

[9] Kshetri, N., & Voas, J. (2019). Big data, dark web and cybersecurity intelligence. Journal of Cybersecurity, 5(1), tyy012.

[10] Piekarski, D., & Piekarski, M. (2019). Web Scraping in Cybersecurity: Scenarios, Opportunities and Challenges. In 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA) (pp. 1-7). IEEE.

[11] Zawoad, S., Hasan, R., & Hasan, K. M. (2015). Dark web crawling: a case study of Jihadist forums. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 384-389). ACM.

[12] Zhu, H., & Zhang, R. (2019). Detecting cybersecurity events with deep learning models based on dark web forum data. Computers & Security, 82, 284-296.

[13] "Scraping the Dark Web: A Framework for Large-Scale Online Data Collection" by Mohamed Ali Kaafar, Roksana Boreli, and Vern Paxson (2015)

[14] "Web Scraping: A Review of Techniques and Tools" by Kavita Ganesan, ChengXiang Zhai, and Jiawei Han (2009)

[15] "Scraping Intelligence: A Framework for Web Intelligence Gathering" by Alaaeldin M. Hafez and Reda A. El-Bashir (2018)

[16] Shan, Z., Zhang, Y., Yang, Q., & Xie, H. (2012). Mining security discussion forums for zero-day exploit prevention. Expert Systems with Applications, 39(1), 473-481.

[17] Wang, J., Wu, X., Lu, Y., & Zhu, X. (2011). Automatic identification of web vulnerabilities from security discussion forums. In Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics (pp. 217-222)