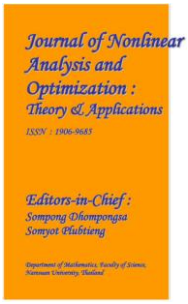
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2, No. 4: 2023
ISSN : **1906-9685**



TECHNIQUES TO PREVENT ATTACKS CAUSED THROUGH WEB SCRAPING - A REVIEW APPROACH

Dr. Hannah Vijaykumar Associate Professor & Head, Department of Computer Science Anna Adarsh College for Women, Chennai, Tamilnadu, India hannahvijaykumar@annaadarsh.edu.in

ABSTRACT
As the technology is growing now a day, Consumers have to understand how they may use old content accessible and impact their environments, since technology keeps advancing and information is still seen as the top priority. These days, data can be unorganized, enhanced, unprocessed, arranged or further refined. Data always contributes to the growth of all fields, and its analysis is necessary to keep up with demand. Web scraping has become crucial for the development of enterprises, scientific research, and even for obtaining information about the most popular online articles. Web scraping has grown crucial for the expansion of enterprises, scientific research, and even for determining which articles are the most popular online. In web scraping, information stored in a database is extracted along with the supporting HTML code. After that, the web scraper can duplicate a whole website's information offsite. Web scraping can quickly result in more serious attacks. Web scraping is a common practice at many organizations as one of the first steps in teams or ransom ware activity. This Paper describes a review about types of scraping techniques and attacks caused by using web Scrapers.

Keywords: Web Scraping, Information, HTML Code, Website, Attacks, Web scraping techniques.

1. INTRODUCTION

The extraction of information and material via an online resource is known as web scraping. A great deal of individuals with computers engage with the internet via visiting websites through browsers, where data and multimedia items are presented in a user-friendly way for simple comprehension. In terms of the infinite possibilities of the web, this is just the surface of the iceberg because there may be so much valuable unprocessed data that is obscured from view. Web scrapers are utilized by many digital enterprises, including search engine providers, price comparison sites, and market research firms, for legal as well as illegal purposes. Web scraping has numerous acceptable uses, such when a search engine robot crawls a page, analyses its information, and then ranks it, when Websites that compare prices use Bots to automatically extract pricing and product details from dealer web pages. The various websites that fall under the given category present information in a variety of ways. You might not be able to view all the info at once even using a single website [1]. The information may be spread out over several pages in different areas. Web scraping API is a web scraping service that is more extensive and based on specific queries. By using the API of view, it gives institutions access to client data that has been scraped and converted into structured data. Web scraping is mostly used for content scraping, which involves obtaining the website's content for purposes including mining data, knowledge indexing and price scratching which extracts comparable costs for online analysis, web mashup, and integration of data. With the exponential growth of Internet users, there has been a huge rise in the quantity of online enterprises, from ticketing and job portals to online shopping and online content creation. Online businesses make sense of their Web traffic, the majority of it will come from bots rather than real people. Furthermore, they can maintain their competitive advantage if Web scraping bots are made with nefarious intentions. Searching for indications of Botnet usage is one

method of spotting malicious web scraping activities. Due to the immense quantity of information gathered by each site, web scraper bots require an enormous number of resources to operate.

2. LITERATURE SURVEY

Web scraping is often referred to as screen scraping, web harvesting, web data extraction, and web scraping. Data mining techniques include web scraping. Web scraping is the process of extracting information from web pages and converting it into a format that can be understood, such as spreadsheets, databases. The most important issue that needs to be resolved in order to realize the semantic web is the extraction of relevant information from the web. Semantic annotating, web usage mining, and web scraping are just a few of the methods that can be used to accomplish this. Web mining makes it possible to locate pertinent results on the internet and is used to draw out important data from the discovery patterns stored on the servers. Web utilization mining is an instance of web analysis that gathers data on users' access methods and browsing patterns. Different formats are used to store various types of data. The internet has an enormous amount of website content and data sources available in a variety of forms, including print, multitasking, recordings, and footage, which are going to be determined by the inconsistent nature of restitution due to the triviality of the client's observation that the information displayed on sites is accurate. This page has been visited via a web browser. They lack the necessary features. The only option left is to manually copy and paste the information from the internet source into a plain-text file in the internet browser and then save it to computers hard disc. The web scraping approach has been used to create an effective application for gathering data from Instagram profiles. Application testing was carried out among 15 accounts in study, with an average of 100–11,000 articles. Researchers can locate relevant publications for further analysis by using web scraping for web analysis for citation. Web Citations based on a specific query can be done using the Hidden Markov Model, Firefly Optimization and Particle Swarm Optimization algorithms. The sole basis of healthcare is no more physical interaction. Instead, it has gone digital in its own special way. Web scraping for health purposes can save many lives by enabling informed judgments in this data-driven society [2].

3. WEB SCRAPING TECHNIQUES

Screen scraping is a technique for collecting data that is used to scrape data from displays for later usage. Screen scraping is typically used to gather information from a single programme and transfer it to another, or, more controversially, to take away information. Onscreen information, such as text, photos, or charts that display on the desktop, in an application, or on a website, is captured as plain text and converted into visual data. Screen scraping can be done manually by a person gathering data or dynamically by a Scraping Programme. There are several methods used in Web Scraping:

- Human Copy & Paste
- HTML Parsing
- HTTP Programming
- DOM Parsing
- Semantic Annotation Recognizing
- Computer Vision Webpage Analysis

3.1 HUMAN COPY & PASTE

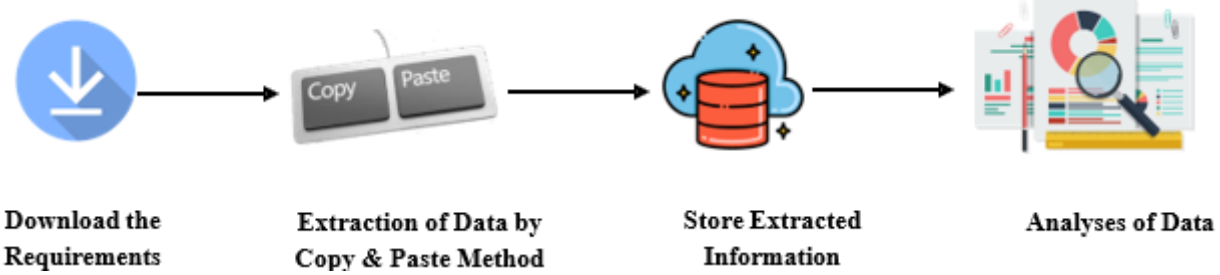


Figure 1: Work Flow of Human Copy & Paste for Web Scraping

The above mentioned figure 1, states the flow of Copy & Paste mechanism of web Scraping. The process of copying and pasting the required information is straightforward approach: open the website in the browser, and then actively duplicate it by pasting it onto another piece of media. The method is rather simple and quick to use, but if the website uses a barrier programme, it makes it tough to use because it necessitates a manual selection of somewhat lengthy objects or statements. However, other techniques require an additional programme and are more difficult to use. Manual web scraping enables individuals to verify of each data point to prevent inconsistencies or exclude inappropriate data from extraction. Because time is increasingly seen as being more precious, manual web scraping can be rather expensive, in terms of time.

3.2 HTML PARSING

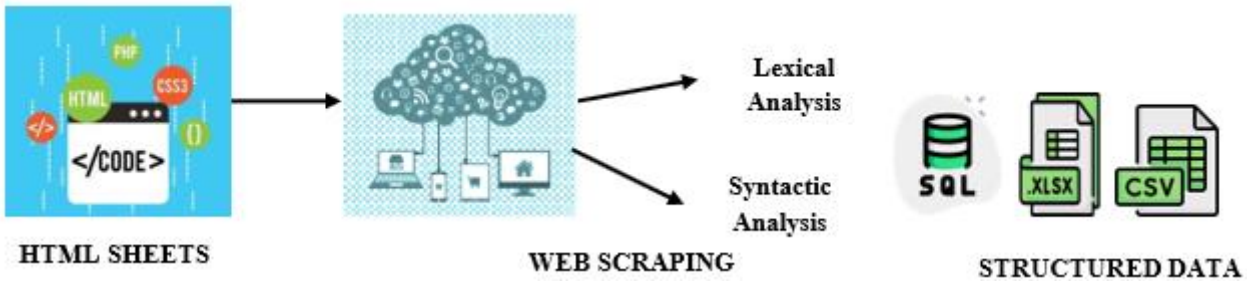


Figure 2: Web Scraping based on HTML Parsing

The above figure 2, States that the Web Scraping using HTML Parsing, which operates in two ways: Lexical analysis and Syntactic analysis. In data mining, a wrapper is a programme that recognizes templates in a certain source of information that retrieves its content, and transforms the information into a structured form. In several websites, huge numbers of Web Pages are generated systematically from a basic structured source, such as database. Content representing the identical area is encoded into related pages using a shared script or template. The language used to create the web scraper can be of any form, the first step is to submit a request to the servers to acquire an unprocessed HTML content in order to obtain permission to view the data on a website. This data formatted in HTML is essentially incomprehensible [10]. Parsing information using HTML is the process of turning an unorganized stream of data into a tree or parse tree, which is simpler to interpret, understand, and apply. HTML is the most common format in which web scraping data is delivered. Because of this, there are numerous freely available HTML parsing packages accessible for practically all languages.

3.3 HTTP PROGRAMMING

Web Scraping Method based on HTTP Programming is considered as a less common technique and more creative, to scrape by sending direct HTTP queries to the HTTP endpoints that run the programme or website. Links to a connecting point known as an HTTP endpoint that exposes data, HTML files, or live server pages. One can use the Python module to send HTTP requests to obtain contents. By submitting Receive or Send requests, it is possible to access the HTML of a website or it's API. A straightforward HTTP request is all that is necessary to extract content from a Website and then save it as plain text.

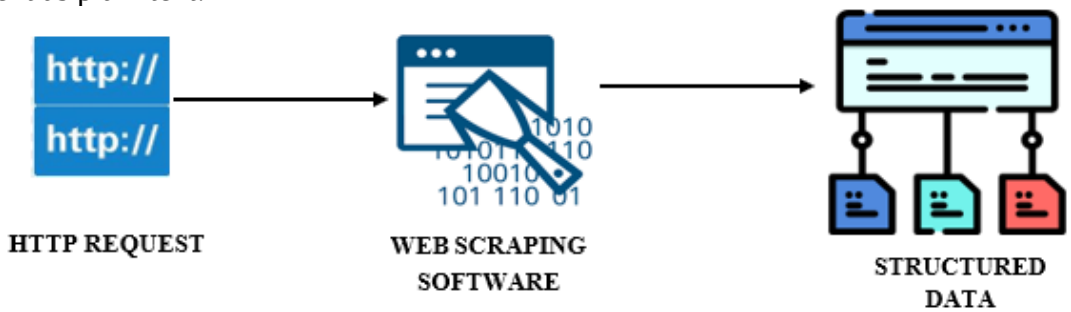


Figure 3 : WEB SCRAPING USING HTTP PROGRAMMING

3.4 DOM PARSING

Targeting the chosen DOM (Document Object Model) elements of the website and then analyzing or saving the content contained within those DOM elements constitutes web scraping. A PHP API that parses the entire page and searches the DOM for the necessary items can accomplish the same thing. For more complicated HTML pages, this is typically quite time-consuming, and it frequently fails when the HTML DOM marginally changes. The below described figure 4 states the web scraping using DOM Parsing [9]. It doesn't provide the ability to use an earlier version of this data that has been already saved. Therefore, the only choice is to manually copy and paste the necessary data, which is actually a very difficult task and may take hours to finish. After downloading the HTML DOM from a website server, it is necessary to parse the DOM tree and extract the necessary data. After downloading the HTML DOM from a website server, it is necessary to parse the DOM tree and extract the necessary data. Articles retrieved from Google Scholar are extracted by HTML DOM Parser using PHP. Information from a Research article includes the title, authors, citations, links, and the year of publication [3]. Finally, MYSQL Server is used to store the scraped data for later analysis. Articles retrieved from Google Scholar are extracted by HTML DOM Parser using PHP.

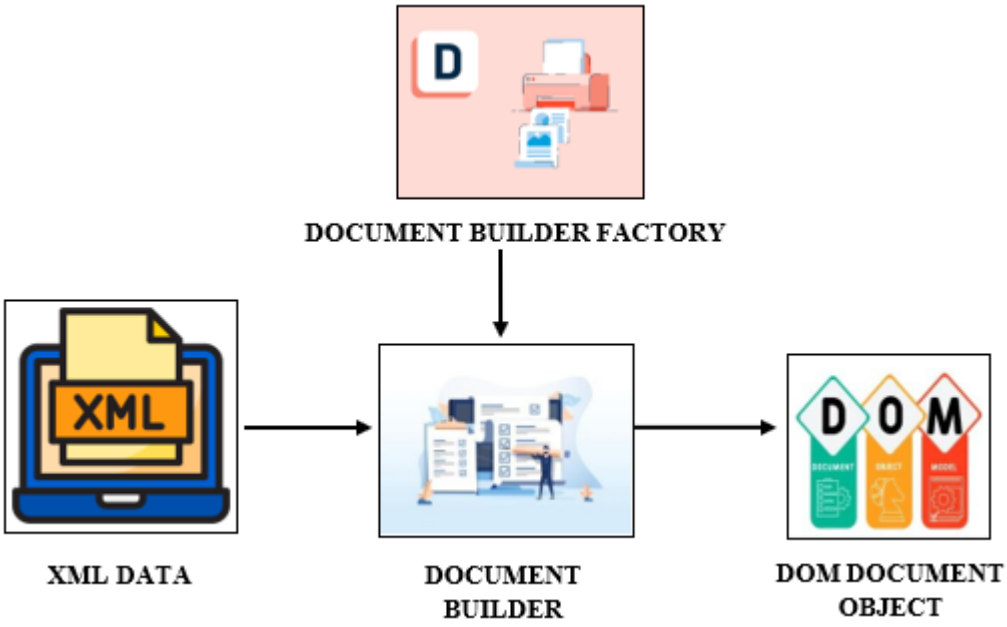
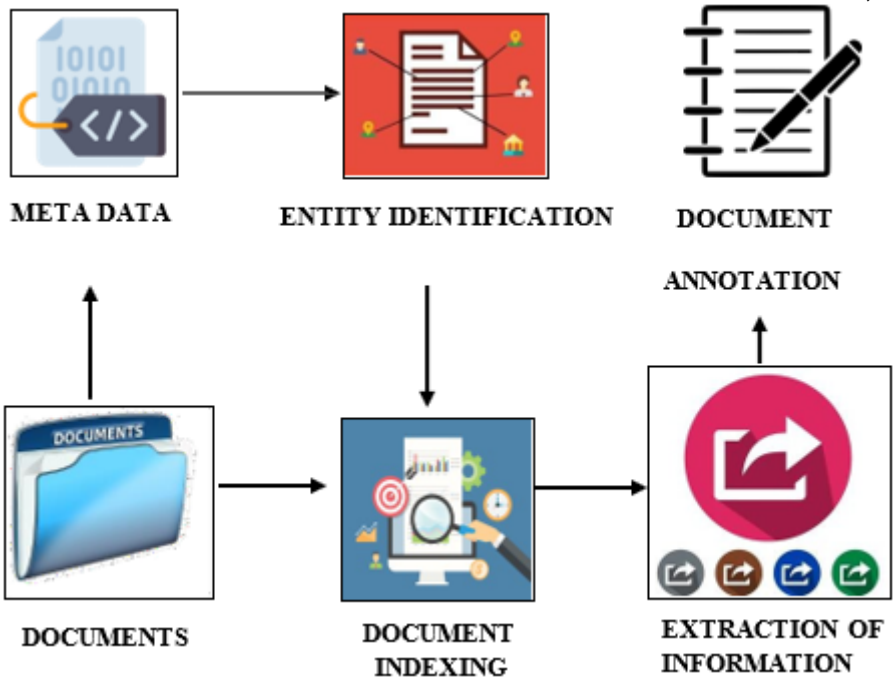


FIGURE 4: WEB SCRAPING USING DOM PARSING

3.5 SEMANTIC ANNOTATION RECOGNIZING

The goal of semantic annotation, on the other hand, is to make web material more meaningful by using semantic tags or information. By providing structured data about the content, these annotations help machines comprehend and interpret the material. In order to mark up web pages with semantic information, semantic annotation approaches include the use of RDF (Resource Description Framework) triples, microdata, or JSON-LD (JavaScript Object Notation for Linked Data). Web scraping can be used to retrieve data from webpages without the need for semantic annotations, but combining it with these approaches might have even more advantages. You can improve data extraction precision, enable better data integration, or facilitate compatibility with other systems that understand semantic annotations by recognising and extracting semantically annotated data [4]. Overall, web scraping and semantic annotation approaches can be used to provide more accurate and efficient data extraction from websites, especially when working with structured and semantically marked-up content. The below figure 5,states the web scraping method using semantic annotation recognizing.



5: WEB SCRAPING USING SEMANTIC ANNOTATION RECOGNIZING

3.6 COMPUTER VISION WEBPAGE ANALYSIS

Using image processing and machine learning methods, web scraping with computer vision approaches entails extracting data from online pages based on their visual content. Computer Vision web scraping focuses on analyzing and comprehending the visual components of a web page rather than using the conventional ways of parsing HTML or accessing DOM elements. It's crucial to remember that using computer vision to scrape the web has significant restrictions and difficulties [5]. Algorithms for computer vision may have trouble with websites that have intricate layouts, dynamic content, or a lot of JavaScript usage. Furthermore, because computer vision techniques depend on visual cues, any modifications to the visual design or organizational framework of the web page can necessitate modifying the scraping algorithm. Overall, when dealing with websites where conventional scraping methods are less successful, web scraping utilizing computer vision can be a valuable strategy. The below mentioned figure 6 describes the computer vision webpage analysis based web scraping to build an efficient algorithms for extracting data from web pages based on their visual content, it necessitates competence in computer vision, image processing, and machine learning.

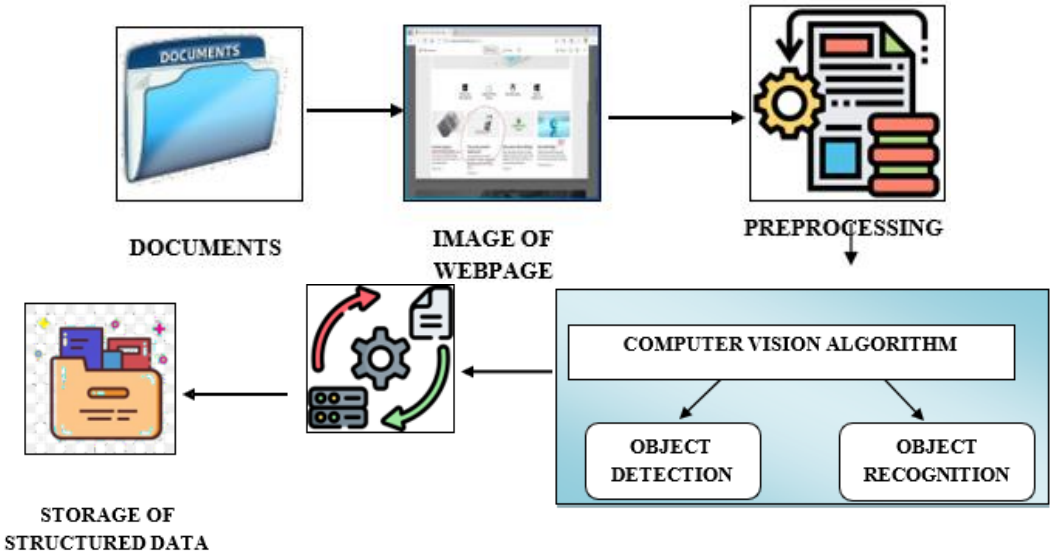


FIGURE 6: WEB PAGE ANALYSIS THROUGH COMPUTER VISION

Web scraping is a method for getting information from WebPages; it is not inherently dangerous. Web Scraping can, be used improperly or without authorization, which can result in a number of negative outcomes and possible attacks.

3.7 DENIAL OF SERVICE (DOS): DoS can occur when a website is aggressively and excessively scraped, which results in a huge rise in server load and a Denial of Service attack. The volume of requests may make it impossible for authorized users to access the website [6]. Web scraping alone may lead to a Denial of Service (DoS) attack in some situations. A case in point is as follows:

3.7.1 Targeted Scraping: A hacker chooses a website to scrape and steal information from, frequently with harmful intent. To automate the scraping procedure, the attacker may utilize online scraping tools or unique scripts.

3.7.2 Excessive Requests: The attacker initiates a broad-based scraping operation and quickly floods the target website with excessive requests. To spread out the burden and make it more difficult to block their requests, they might employ several IP addresses, proxies, or Botnet.

3.7.3 Impact on Availability: The additional server workload brought on by the scraping activity may cause the target website to become temporarily or permanently unavailable, preventing legitimate visitors from accessing it. This service disruption could have a detrimental effect on companies who rely on the website, disrupt online services, and ruin the user experience.

3.8 CONTENT THEFT AND COPYRIGHT INFRINGEMENT: Copyright infringement and intellectual property breaches can result from downloading content without authorization from websites. Some people scrape content in order to post it elsewhere under their own names or to build replica websites for nefarious purposes [8].

3.8.1. Unauthorized Content Extraction: Without the proper consent of the website owner or content creator, an attacker may scrape a website's content, including text, photos, videos, or any other media. This scraping procedure could entail mechanically removing substantial amounts of content from a number of pages or the entire website.

3.8.2 Copyright Violation: Scraping and republishing content without authorization is a copyright violation that violates the owner's exclusive rights. Text, photos, music, videos, and other creative works of authorship are all protected by copyright laws. Unauthorized scraping and usage of protected content may result in litigation, fines, and other legal repercussions for the offender.

3.9 MALICIOUS DATA INJECTION: Web scraping could be used by attackers to insert harmful data or code into a website. They can damage the integrity of the scraped page and potentially infect visitors with malware by taking advantage of flaws in the scraping procedure or introducing malicious payloads.

3.9.1 Exploiting Input Fields: The attackers may interact with input fields, forms, or data submission methods on the target Website while doing the scraping activity. They take advantage of the website's inability to adequately validate user inputs by submitting designed or malicious data.

3.9.2 Payload Injection: The target website's inputs or data endpoints are injected with specially designed data or payloads by the attacker. Cross-site scripting (XSS), command injection, and other injection-based assaults are examples of this [7]. The intention is to take use of the systems, databases, or other elements on the website that handle the data that has been scrapped.

3.10 SCRAPING OF SENSITIVE INFORMATION: There are some Websites may contain sensitive data that should not be viewed or collected without proper authorization. Attackers could cause serious harm to people or organizations if they scrape personal data like financial records, medical information, or trade secrets.

3.10.1 Data Breaches: There is a chance that an online scraper could unintentionally expose or improperly manage sensitive data, such as personally identifiable information (PII), financial information, or medical records, during a web scraping process. This could result in a data breach. A compromise like this one could lead to identity theft or unauthorized access to sensitive data.

3.10.2 Web Scraping Detection Evasion: Some websites use a variety of countermeasures to identify and prevent scraping operations. Attackers may employ methods like IP rotation, user agent spoofing, or distributed scraping infrastructures to try to get past detection systems [7]. The resources of the

target website may be put under more stress as a result of these evasion techniques, or website administrators may even take corrective action.
When performing online scraping activities, it is essential to follow moral and legal rules in order to reduce these dangers and prevent potential assaults. It is advised to obtain the appropriate consent, abide by the terms of service of the website, take precautions to secure personal information, and keep in mind how sensitive the information being scraped is. Additionally, using the right encryption, anonymization, and access control methods can help safeguard sensitive data while it is being stored and transmitted.

4. CONCLUSION

In conclusion, web scraping is a useful method for mechanically obtaining data from websites. It entails locating and extracting specified data by parsing the HTML or XML structure of online pages using a variety of tools and techniques. Tools like Selenium can automate web browsers to render and interact with the page while dealing with dynamic information before extracting data. While online scraping, it's crucial to be aware of website policies and to observe their terms of service. Web scraping also involves handling pagination, using anti-scraping precautions, and being aware of the moral and legal ramifications. Overall, online scraping techniques provide effective ways to collect data from websites for a variety of uses, including research, data analysis, and application development. But it's important to be aware of website regulations, abide by the law, and make ethical use of data that has been scrapped.

REFERENCES

[1] Prof. Shivsagar Gondil, SmitPatne, Tejas Raut, Vinit Bhagat, (2021),"A Survey on Web Scraping and its Applications", International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org pp.43118-4321.
[2] Sakr, S., & Elgammal A. (2016). "Towards a comprehensive data analytics framework for smart healthcare services. Big Data Research, 4, 44-58."
[3] Lotfi, C., Srinivasan, S., Ertz, M., Latrous, I. (2023). A Tool for Study on Impact of Big Data Technologies on Firm Performance. In: Rajakumar, G., Du, KL., Vuppalapati, C., Beligiannis, G.N. (eds) Intelligent Communication Technologies and Virtual Mobile Networks. Lecture Notes on Data Engineering and Communications Technologies, vol 131. Springer, Singapore. https://doi.org/10.1007/978-981-19-1844- 5_40
[4] Sirisuriya, De S. "A comparative study on web scraping." (2015) INTERNATIONAL RESEARCH CONFERENCE ARTICLES (KDU IRC) http://ir.kdu.ac.lk/handle/345/1051.
[5] P. Marques et al., "Detecting Malicious Web Scraping Activity: A Study with Diverse Detectors," 2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC), Taipei, Taiwan, 2018, pp. 269-278, doi: 10.1109/PRDC.2018.00049.
[6] Chao-Yang, Z. (2011, August). DOS attack analysis and study of new measures to prevent. In 2011 International Conference on Intelligence Science and Information Engineering (pp. 426-429). IEEE.
[7] Parikh, K., Singh, D., Yadav, D., & Rathod, M. (2018). Detection of web scraping using machine learning. Open access international journal of Science and Engineering, 114-118.
[8] Potluri, V., Potluri, S. S., Tummala, G., & Bolla, S. (2021). Online Copyright Infringement. International Journal of Engineering Research & Technology (IJERT), 10(3), 127-133.
[9] Salem, H., & Mazzara, M. (2020, December). Pattern Matching-based scraping of news websites. In Journal of Physics: Conference Series (Vol. 1694, No. 1, p. 012011). IOP Publishing.
[10] Khder, Moaiad. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and its Applications. 13. 145-168. 10.15849/IJASCA.211128.11.