International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Web Scraping and It's Application: State of the Art, Technique, A Review.

Nehemiah Stephen
Department of Computer Science & I.T
V.I.A.E.T, Prayagraj, India
nehemiah_stephen@yahoo.com

Vineesh Cutting
Department of Computer Science & I.T
V.I.A.E.T, Prayagraj, India
vineesh2pro@yahoo.co.in

Abstract: Web scraping is basically an interactive method for website and some other online sources to browse for and access data. To delete a replica of the information and save it in an external archive for review, it uses software engineering technology and custom software programming to extract data or any other content of on-line sources. Web scraping is often called automatic data gathering, database discovery, database crawling, or content management mining. Web scraping have possibly existed since before the start of the World Wide Web, but it has been used mainly in the context of data analytics, and is generally associated to e-commerce. Web scraping technique provides a broad collection of options and can serve various purposes: A web crawler's least necessity is to automate the normally physical work of gathering cost quotation marks and website article details. A web crawler's main requirement will be to discover formerly inaccessible sources of price data, and include a survey of all accessible price information. This scraping process is performed using different technologies which can be automatic application tools or manual methods. This paper provides the overall review of web scraping technology, how it is carried out and the effects of this technology.

Keywords: Web scraping, E-commerce, Data extraction, Web crawler, automatic tools.

1. Introduction:

Web scraping is not initially developed for research of social science, as a effect, analysts using this method may incorporate unknown suppositions into their own, because web scraping will not usually require direct contact among the analyst and those who were formerly collecting the information and inserting it online, data analysis issues may simply arise. Research teams using web scraping techniques as an information gathering method still have to be acquainted with the accuracy and correct analysis of the details retrieved from the website. One final problem analysts must address is the potential effect of web scraping on a publication's functionality, as certain web scraping actions unintentionally overpowered and close down a webpage. A web scraper which is appropriately intended and executed, could assist analysts prevail over obstacle to data access, gather online information more resourcefully, and eventually respond investigation queries that cannot be answered by conventional means of assortment and examination. The below figure 1 shows the overview of how web scraping is done.

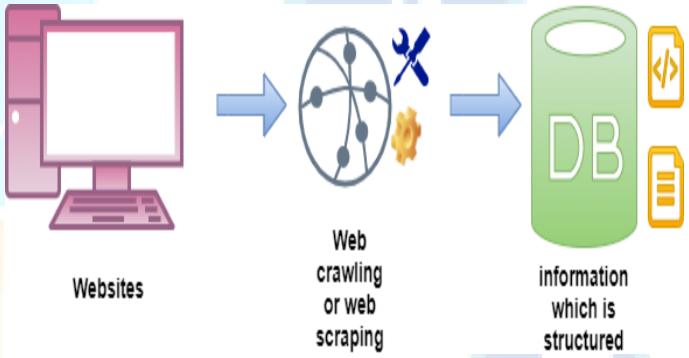

Fig. 1: Web scraping technology

Web scraping API is the framework that allows a company to expand its current internetbased infrastructure, as well as the collection of services designed to build new platforms, integrate designers or integrate collaborators. It allows delivering clean and organized data from current websites, such that different systems can access the data easily. Data disclosed over these APIs could be easily monitored, converted and managed. The underlying architecture lets developers implement improvements to the website without disrupting the structure of the abstraction by shifting them to modifications. Web Scraping Technology is the automated system used to necessarily make the automatic copying and pasting work. It often gathers vast quantities of data from websites such as directory pages, property investment pages, confidential pages and job sites and is stored on multiple servers. Web scraping could algorithmically optimize physical work by accessing each page, extracting information from websites, and html page decoding. There is also a range of Web Scraping Software on the marketplace which can enable to scrap information from every page you need.

2. Related Work:

Renita Crystal Pereira et. al., [1] provided web scraping summary and techniques and tools that face several complexities as data extraction isn't that simple. These strategies guarantee that the data collected is correct, consistent and has better integrity, because there is a large

amount of data present which is hard to handle and retain. Although there are a few problems faced by functional techniques that can be such as the elevated amount of web scraping be able to cause rigid harm to the websites. The measurement level of the web scraper will vary with the measurement units of the original source file, making it very difficult to interpret the data.

Using social networking sites and internet is amplifying day by day like facebook, twitter, linked-in and some other, user knowledge is also high in the internet available from everywhere. This as well offers hackers an advantage in stealing information. Where the concept of rising income comes into being, social networking is important from a view of business point. Like with online shopping, it will also assist consumers in getting fast shopping and also save time. On the other hand, there is advantage in supporting the company and profiting from it.

Kaushal Parikh et. al., [2] proposed a web scraping detection with the help of machine learning It is valuable for research dependent companies. Web scraping has forever been a difficult preventive attack. Every time a company places its data on internet, it is probable that it could be copied and pasted and then utilized in the other point of view without the corporation knowing itself about it. A lot of protection mechanisms have already been in place but some of them continue to be ignored. The significance of machine learning therefore steps in. Machine learning is quite effective on pattern detection. Therefore if we succeed in making the machine understand a cadence of intruder then it will avoid these types of threats from occurring.

Web scraping solutions are aimed primarily at translating complex data obtained through networks into structured data that could be stored and examined in a central database. Web scraping solutions thus have a significant impact on the result of the cause.

Sameer Padghan et. al., [3] projected an approach where data extraction is done from web pages in assistance with web scraping easily. This method would enable the data to be scrapped from numerous websites that will minimize human intervention, save time and also enhance the quality of data relevance. It will also support the user in gathering data from the site and to save the data to their intent and use it as the individual wishes. The scraped information may be used for database development or for research purposes and also for different similar activities. The scraping used would increase significantly and will often encroach on the framework to obtain the details. However the scraping can be stopped by using effective and safe-web scraping methods. This method should be treated as a blessing that must be used carefully for the advancement of human races.

Anand Saurkar et. al., [4] discovered latest technique named Web Scraping. Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structured data, subsequently collected and evaluated in spreadsheets in central database. This research focuses on a summary of the data extraction process of web scraping, various web scraping strategies and most of the latest tools utilized to scrap web. The primary function of this methodology has been to get webbased information and integrate this into a specific repository. The authors addressed the basics of Web processing in this article. They concentrated on the Web scraping techniques. The final part of the paper presents a summary of the numerous technological resources that are available for effective web scraping in the industry.

Federico Polidoro et. al., [5] concentrated on the outcomes of web scraping evaluation strategies with particular orientation to user electronics services and goods throughout the sector of commodity price studies. Although the research done has so far been performed in a small amount of time, that you can see in whatever followed, it has enabled to attain important, but not conclusive, novel efficiencies results.

Web scraping strategies used in the growth analysis will provide exposure to a greater volume of data than that accessible in the existing data set, thus, with the potential to increase the growth estimate. This topic has been briefly addressed in the portions allocated to both of the examined items, but in reality interacting with this viewpoint requires a concern regarding the current survey architecture that does not require or only selectively permit the use of big data approaches within the existing sampling frameworks.

Jan Kinne et. al., [6] Proposed a web extraction platform for the accurate and measurable mining of ecosystems for development. Researchers have put special emphasis on exploring a possible bias while examining technology structures across corporation website if all those types of companies could be measured using suggested method. Web extraction still has to deal with incredibly large and ultra-connected outer websites as a research tool, and the reality that limited broadband access continues to discourage companies from managing their internal websites and therefore preventing themselves from web mining research. The proposed system of research enables for an integrated, least expensive simulation of whole business communities, that could be conducted out more efficiently and in relatively short time periods compared to conventional techniques. This method is also conveniently extendable by checking the web pages of research institutions to model information communities. The key point in proposed system is to identify and extract certain bits of data from unstructured content on the site which exposes information regarding the current development practices of companies.

Ingolf Boettcher [7] discovered that technique like web scraping can evolve. Web scraping innovation provides a range of choices and can satisfy various purposes: A web crawler's basic requirement is to automate the normally physical work of gathering price estimates and website article details. A web crawler's ultimate requirement will be to discover previously inaccessible pricing data outlets and include a census of all web-available price information. The actions to build web scraping for price analytics include significant analytical and administrative consequences. Any deployment of the approach involves a detailed preparation in various sectors. Elements of ethical and data protection need

to be discussed first. Essential IT services and IT practice needed to manage the automated information gathering system must be calculated and must not be overlooked in the context of a research project.

Erin Farley et. al., [8] destined to present web scraping to law enforcement researchers and illustrate what web scraping is about and how this technique works. Use of the web crawling by investigators in criminal justice is a fairly recent trend. Only a range of experiments wherein web scraping was used were identified in a literature review for criminal libertyrelated research using web scraping as an information gathering method. Although web scraping is usually seen as a method for collection of data to promote analysis and research, designing and implementing a web scraper includes technological abilities that researchers in the social sciences generally do not have. A strong level of expertise in computer science techniques like R or Python when developing source code is a necessity for creating a web scraper.

The popularity of the Internet led to rapid increase in amount of data created each day, which prompted the need to scrape websites. Search engine engineers created the first well recognized scrapers like Google. These scrapers search over the whole Internet, scanning each web page, extracting information, and compiling a searchable index. (Lawson, 2015)

Web scraping/web crawling is extracting information from website through the computer software, this software can will copy the way the human explore the world wide web, by executing a fully-fledged web browser, like Mozilla Firefox or internet explorer, also another way to copy the way the human the world wide web, it through the low-level Hypertext Transfer Protocol (HTTP). In other words, instead of copy-paste information manually from the website and gathering it in a spreadsheet, web scraping offers this functionality in computer applications that can accomplish it more precisely and much quicker than a human. (Lawson, 2015)

Web scraping is a technique for converting unstructured web data into structured data that can be saved and analyzed in a central spreadsheet or database. This enables the bot to retrieve large volumes of data in short amount of time, which is advantageous in today world especially since we have big data which is always changing and updating. For example, assume you have a clothing shop and you want to keep track of competitor cloth pricing. You can go to competitor website and gather information every day in order to compare them with your pricing for each product, but this will take a lot of time. Furthermore, if you have thousands of competitors it will be very difficult to keep track of the pricing this is where web scraping comes in handy. (Banerjee, 2014)

Web scraping is a technique for converting unstructured web data into structured data that can be stored and analyzed in a central database or spreadsheet (Sirisuriya, 2015). Web scraping is used in different industries such as cyber security, cyber security, etc. it can be used to determine prices and costs of production by assessing the data that has been gathered. Ads can be created to advertise products to different users depending on their data such as location and cookie settings. It can be relative described as a new method for data collection on the internet. Web scraping is already being used for scientific and commercial applications since it allows for the development of big, customized data sets at minimal costs. It is meant to replace outdated methods of data collection as more business are looking to catch up on new trends that are followed by consumers. (Hillen, 2019).

Web scraping API which is a web scraping service that is in larger scale, based on customized requests web scraping. It provides institutions with structured data by accessing to scraped data of their clients using API. (Mitchell, 2018)

Web scraping was the only way for a program to obtain information from the internet until the recent emergence of APIs, which are special interface designed to make communication easier for applications and servers. API are incredible tools which can provide the data in an organized way. API can be access for variety of different data types like Wikipedia article, tweets in twitter and many more. Even though some website offers API, but not all API are free to use and also, they are very limited in terms of how much data they provide and what data they provide. Furthermore, website developer's priority is to focus and maintain the frontend interface over backend API. To summarize, we cannot depend on APIs for acquire the internet data we need, so we need to apply web scraping strategies in order to ensure that the data we are getting matches with what the business or company needs. (Mitchell, 2018) (Broucke and Baesens, 2018).

2.1 Web Scraping Concepts and Techniques

Data is extracted from sites using the HTTP protocol used by web browsers. This process can be done with manual browsing or automated with web crawlers. Webs scrapping is one the most valuable tools available to data scientist as it allows the extraction of huge amount of data that is constantly generated online with a relatively low cost. (Zhao, 2017). Proper data preprocessing and cleanup is required for proper utilization of scraped datasets (Tarannum, 2019) discuss cleanup methods for scrapped data in python. (Manjushree and Sharvani, 2020) Perform a review of web scrapping technology from literature and discuss tools, methodology and process for web scrapping they also discuss the fields and targets of web scrapping. (Grasso et al, 2013) Introduce a new tool and language for web scrapping called OXPath, OXPath is an extension of the XPATH standard which adds several features such as user actions such as clicks and filtering by visual features exposed from CSS such as color. They also discuss several projects utilizing OXPATH such as DIADEM an unsupervised data extraction tool. (Gheorghe et al, 2018) discuss modern techniques and challenges in web scrapping such as captcha, rate limiting, and they discuss a case study of scrapping public transportation data from Bucharest Public Transportation authority website to analyze issues in public transportation.

Traditional copy and paste technique: The most effective and practical web scraping technique is usually a copy-and-paste

and manual analysis technique. However, when users need to scrap a large number of datasets, this is a mistake, tedious, and unpleasant procedure. (Sirisuriya, 2015)

Grabbing text and using regular expressions: This is a significant and simple method for extracting data from web pages. This strategy is based on the UNIX command or the computer language's regular expression-matching capabilities. (Sirisuriya, 2015)

Programming with the HTTP: Users can access data in both static and dynamic webpages applying this method. Data may be retrieved using socket programming via making HTTP requests to a remote web server. (Sirisuriya, 2015)

HTML parsing: semi-structured data query languages is used for parsing webpages HTML code and retrieving and transforming page content

DOM parsing: embedding a full-featured web browser, like: Mozilla browser or Internet Explorer control, programs may access dynamic material created by client-side scripts. These browser controls also parse webpages into a Document Object Model tree, from which applications may get some pages contents. (Saurkar et al, 2018)

Web Scraping Software: Nowadays several tools that can provide a custom web scraping. These programs can identify page data structure automatically or give a recording interface which eliminates the need for web scraping scripts. Furthermore, some of these software's can have a scripting function which can be used for extracting and transforming material, as well as database interfaces for scraping data and storing it in local databases. (Saurkar et al, 2018)

Computer vision-based web page analyzers: By visually scanning web pages like a person, computer vision and machine learning are being utilized to discover and retrieve important information. (Saurkar et al, 2018). A very simple illustration for web scraping is shown in fig 2.

2.2 Web Scraping Process

The web scraping process is divided into 3 stages as shown in fig.3, which are:

Fetching stage: The desired website with the relevant information must first be accessed in what is known as the fetching phase this is accomplished via the HTTP protocol, which is an Internet protocol for sending and receiving requests from web servers. Web browsers utilize similar methods to get material on web pages. In this step, libraries such as curl 2 and wget 3 can be used by sending an HTTP GET request to the target address (URL) and get the HTML page as a response (Persson, 2019)

Extraction stage: After retrieving the HTML page, the important data should be extracted. Regular expressions, HTML parsing libraries, and XPath queries are utilized in this step, which is referred to as the extraction stage. The XML Path Language (XPath) is a tool for finding information in documents. This is the second phase of the project. (Persson, 2019)

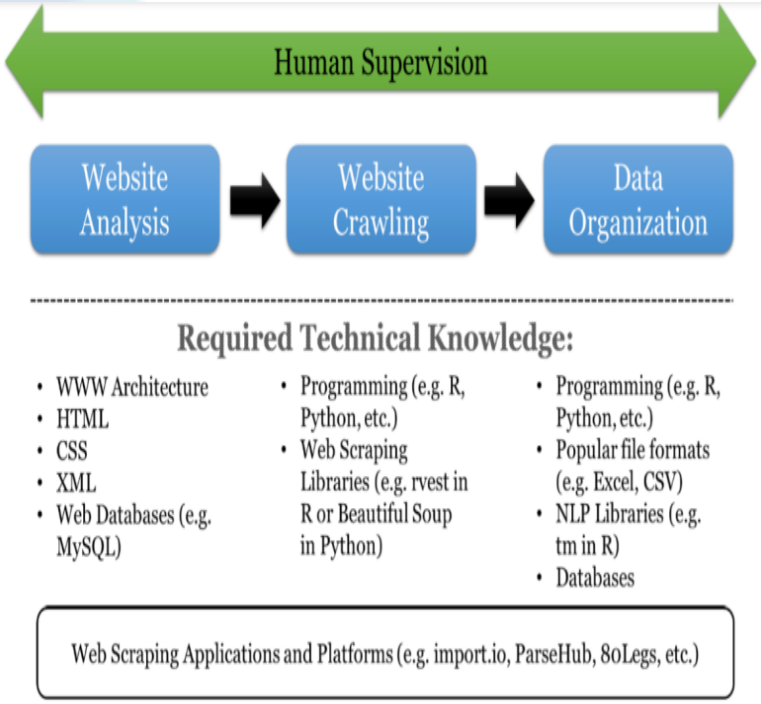

Fig 2: Web Scraping (Krotov and Tennyson 2018)

Transformation stage: Now that just the relevant data remains, it may be converted into a structured format for presentation or storage. Using the stored data, we can gather information which can help the business intelligence in making a better decision and much more. (Persson, 2019)

2.2.1 Python for Web Scraping

One of the reasons why python is a good choice for web scraping is that python is one of the most popular language, some people might not hear about python programming language but it's very easy to learn even if you have not used it before or have not programing experience.

Since python is a popular language it means that if you have a huge community that can help you in case you face any problem or issue, if you have an issue during your programming, it's pretty likely that someone else has had the same problem and solved it somewhere online. This isn't exclusive to Python, but its accessibility and long-standing popularity have resulted in one of the largest and most diversified user communities among today's main programming languages. (CreateSpace Independent Publishing Platform, 2015)

Another reason python is a fantastic choice for web scraping (or learning any programming skill, for that matter) is that Python code is very simple to read. Consider the following example in 3 different programing language. (CreateSpace Independent Publishing Platform, 2015)

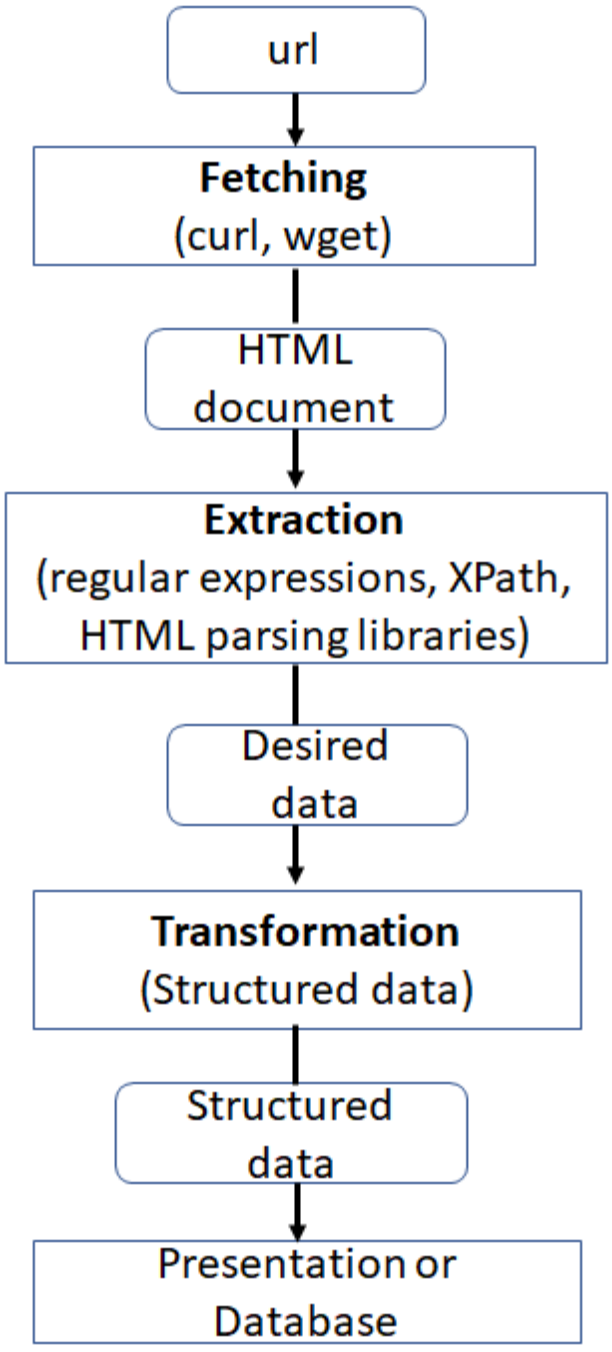

Fig.3: Web Scraping process (Persson, 2019).

Writing a readable code is important not only because it's easier to keep track while you are working on it but also it allows other programmers to understand what you accomplished and it's much easier to figure out what you were thinking when you wrote it after a few weeks or years. This is why well-written Python code is extremely easy to maintain and reuse. (CreateSpace Independent Publishing Platform, 2015). Very simple illustration for the same code written in 3 different languages (Java, C++, and Python) is shown in fig 3, and its noticeable that python is very easy to read.

3. Usage of Web Scrapping

Web scraping techniques allow users to scrape data from various websites into a single database or spreadsheet. As a result, data may be quickly seen and analyzed for future use. (Saurkar et al, 2018)

Web scraping involves the creation and implementation of two software programs: a crawler and a scraper. The crawler downloads data from the Internet in a systematic manner; the scraper then extracts important information in its raw form from the downloaded data, codes it, and stores it in a database or file according to a user-defined structure and format. This new file is then evaluated in ways that the initial data presentation on the internet does not allow for. (Farley and Pierotte, 2017)

Previously, the data was only available on web browsers and it could not be copied. Web scraping has made it possible, and it can be done with a few lines of script in a short amount of time (Parikh et al, 2018)

Web scrapping is widely used in business and scientific fields.(Hillen, 2019) mention the use of web scrapping for food price research due to the low cost of information retrieval, high frequency and details in the generated set compared to other sets such as ones provided by national and commercial agencies, their method involves identifying the target site, selecting the data to be captured and output format and finally automating the process. (Yannikos et al, 2019) discuss the use of web scrapping to extract and analyze darknet marketplaces product supplies and prices, their method involves retrieving the html data through the TOR network, parsing the relevant data into structured form, storage in relational database and finally statistics and charting from stored data. It is used to determine prices of products and expenditure used to spend on advertisements and commercials by businesses. It can recognize trends and user preferences.

4 Application of Web Scraping

Web scraping is used widely in many different computer science fields, especially, the hot trends and news areas such as: Business Intelligence, AI, Data Science, Big Data, Cloud Computing and Cyber Security

4.1 Application of Web Scraping in Business Intelligence

Price tracking, market research, and sentiment analysis are three of the most common uses of web scraping. In e-commerce, price is extremely important. The rivals must be watched from a business standpoint. When prices vary often, manually tracking all rivals' pricing is not a realistic solution. Web scraping serves a function in this situation: it automates the extraction of price information from rivals and can offer

up-to-date information from all of them in one document or database that is easily accessible. (Banerjee, 2014)

For a better decision-making process, market research demands extremely precise data. Market analysts and business intelligence professionals across the world rely on high-quality, meaningful data to do their tasks. This makes online scraping a feasible approach for commercial operations including market pricing, market trend analysis, point of entry optimization, research and development, and competition monitoring. (Vording, 2021)

Web scraping can be used in stock market to visualize the price changes over period of time, and social media comments and feeds have been scraped to know the opinion of the public on opinion leaders. (Sirisuriya, 2015)

Web scraping data from social media for sentiment analysis is a common practice. Predicting elections is one use. By analyzing tweets and postings where the candidate's name isn't even necessary, a computer might guess the winning candidate's name. Sentiment recognition algorithms discover patterns that go beyond the post and detect clues. When utilizing web scraping to collect data, more accurate analysis may be conducted than when using aggregated postings (based on hashtags on Twitter, for example). (Vording, 2021).

Web scraping has a multitude of benefits and advantages. Some of these include:

more efficient, consistent, and easy than manually performing the same tasks, very low maintenance, very low running costs, online reputation Offer your customers more targeted advertising, provides insights into pricing data, market dynamics, prevailing trends, your competitors' practices, and the challenges they face, collect public opinion, scratch lead, develop vertical search engines with a specialized focus, obtaining the most precise and timely results that cannot be obtained by humans and fast and precise findings help firms save time, money, and effort, giving them a clear speed-to-market edge over their competition (Kasereka, 2020; Hillen, 2019; Scraping Expert, 2017)

In addition, the web scraping Have a competitive advantage, where the competitors scrape pricing from e-commerce and marketplace websites using automated bot programs. Competitors use this unlawful behavior to get real-time dynamic pricing data in order to undercut product prices and attract price-sensitive customers. Bot programs may be developed to scrape pricing information from a range of rivals' websites, compare them, and update the portal with the best rates in order to attract more consumers. When potential customers search for items on Google, these rivals come up first, at lower pricing than their competitors. (Krotov and Tennyson,2018; Vording, 2021)

Web scraping is technically possible for every online business that creates original material. The site's susceptibility to scraping is determined by one or more of the following four key factors:

1. Competition from similar businesses is on the horizon.
2. The website's popularity in terms of traffic or user interaction.
3. The stuff that is created/changed is unique or essential in some way.
4. Existing security flaws in websites

However, as a business owner, you can't be sure if you're losing data to the competition or keeping it for educational or research purposes. (Banerjee, 2014; Broucke and Baesens, 2018)

4.2 Applications of Web scraping in AI

Data derived from scientific or academic research is vital in the marketing industry. Scholars, market researchers, and academics gather information from a variety of sources in order to better their products and services. Most website authorities forbid users from downloading data from a website and storing it locally. As a result, the user may desire to manually copy data from the website to a local file storage space on their computer. However, completing such a project would be exceedingly hard and time consuming. Due to these limitations, web scraping methods have been developed. Web scraping techniques allow users to collect data from multiple websites and store it in a single database or spreadsheet. As a result, data may be viewed and evaluated for potential future use with little to no effort. Web scraping is a branch of web mining that focuses on gathering data from the web. An important aim is to provide a comprehensive review of the various web scraping techniques and apps available for the extraction of critical website data. In the Information Retrieval, Data Extraction, and Data Mining triangle, Web Mining is still a vital component. Prior to being processed by data mining tools, retrieval and extraction of important information from unstructured data are critical steps in the process. Exploring these strategies further is critical for dealing with the massive amount of data available in the Information Overload Era. Additionally, as the Web continues to emphasize the importance of semantics and information integration, these fields of study become critical for addressing the Web's new future trends (Saurkar et al, 2018(. The web mining including (data mining, information retrieval and information extraction) is shown in fig 4.

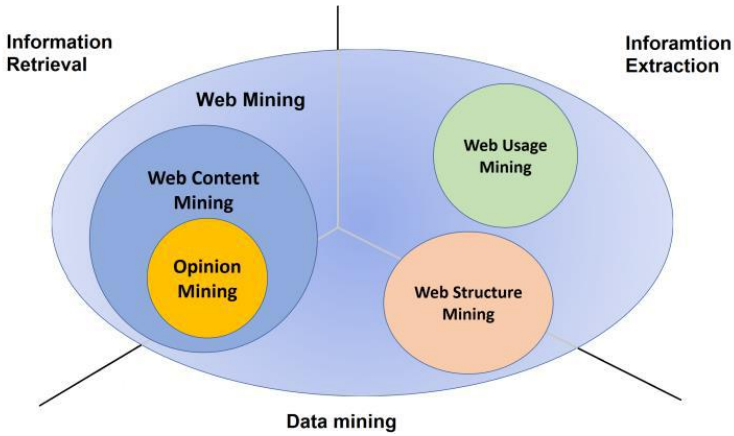

Fig 4: Web Mining (Saurkar et al, 2018)

Information Retrieval (IR) is a topic of study that focuses on the document's retrieval from a collection of various documents (both relevant and irrelevant), typically using key word searches. With the evolution of the Internet, Information Retrieval got a new importance level, as search engines became the leading method of locating data on different websites. Nowadays, due to their prominence, search engines are the top used evocative term in the field of information retrieval. The term "Information Extraction" refers to a sub-discipline of artificial intelligence. Internet Explorer (IE) is mostly concerned with extracting valuable info from unstructured data. Typically, an extraction information system is focused on the identification of individuals or objects (people, places, and businesses, for example) and the extraction of rules. Unstructured data includes video, audio, pictures, and text. Initial information extraction systems concentrated mostly on text, which is remains the most investigated sort of information by commercial frameworks and scientific community to this day. The objective of IE is to extract usable components from unstructured data in order to create more value information via semantic classification. The outcome may be applicable to other information processing tasks, such as information retrieval and data mining. While the objectives of IR and IE are fundamentally different, they should be viewed as closely related activities in order to improve their precision and accuracy.

Data has gained extra value as a result of advancements in data computing and analysis technologies. It was difficult to think several years ago that we would ever be able to pull this much information from the Internet. All of this is possible because modern techniques are capable of processing massive amounts of information in a short time. Internet offers access to a vast volume of unstructured or unlabeled material that is difficult for humans to acquire due to their lack of expertise about available information sources. Additionally, many people are unaware that their personal data is available online. The article discusses a system for retrieving personal information from the Internet based on a variety of input criteria. By utilizing artificial intelligence algorithms, the system is capable of differentiating the information of multiple people with the same name. Although the material for performed case study was acquired from different sources holding data about people in Spain, this may easily be applied to other countries' specialized sources of data. This approach was confirmed in a case study including multiple participants, and the findings were extremely satisfactory. (Chamoso et al, 2020)

In computing, a bot is an autonomous software that operates on a network (particularly the Internet) and can interact with other systems or users. A description of how an Artificial-Intelligence bot's memory can be optimized through the use of a faster searching algorithm and how it can learn new things that the user desires the bot to learn. A Web-Crawler is a type of bot that crawls through a collection of websites or the entire internet. Web-crawlers are also referred to as Web-spiders. While crawling the entire internet is a too much for a personal assistant, a bot can crawl a few sites and gather the information. Therefore, in order for the bot to gather information from the internet, it must crawl through the websites and scrape the necessary information. To accomplish these duties, a web-crawler or spider is utilized for crawling and a web-scraper is used for scraping. (Bhatia, 2016).

5. Conclusion:

Extracting data through scraping technology is a new evolving activity in the technology harvesting arena. Though many companies are still using manual process of extracting data but Web Scraping solutions will transform the traditional method of extracting data. The day is not that far with exponential growth throughout this field when it can become a phenomenon and most companies will understand the value of scraping innovation and how it enables them remain ahead in the race dramatically. This paper presents the survey of Web scraping technology incorporating what it is, how it works, the popular tools and technologies of web scraping, the websites used for this technology and the top most fields which are making use of this technolology.

References

[1] Renita Crystal Pereira and Vanitha T, "Web Scraping of Social Networks," International Journal of Innovative Research in Computer and Communication Engineering, pp. 237-240, Vol. 3, 2015.

[2] Kaushal Parikh, Dilip Singh, Dinesh Yadav and Mansingh Rathod, "Detection of web scraping using machine learning," Open access international journal of Science and Engineering, pp.114-118, Vol. 3, 2018.

[3] Sameer Padghan, Satish Chigle and Rahul Handoo, "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros," Journal of Advances and Scholarly Researches in Allied Education, pp. 691-695, Vol.15, 2018.

[4] Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode, "An Overview On Web Scraping Techniques And Tools," International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 363-367, Vol. 4, 2018.

[5] Federico Polidoro, Riccardo Giannini, Rosanna Lo Conte, Stefano Mosca and Francesca Rossetti, "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation," Statistical Journal of the IAOS, pp. 165-176, 2015.

[6] Jan Kinne and Janna Axenbeck, "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," 2019.

[7] Ingolf Boettcher, "Automatic data collection on the Internet," pp. 1-9, 2015.

[8] Erin J. Farley and Lisa Pierotte, "An Emerging Data Collection Method for Criminal Justice Researchers," Justice Research and statistics association, pp. 1-9, 2017.

[9] CreateSpace Independent Publishing Platform. (2015). Learn Web Scraping with Python in a Day: The Ultimate Crash Course to Learning the Basics of Web Scraping with Python in No Time. CreateSpace Independent Publishing Platform, North Charleston, SC, USA.

[10] Dogucu, M., Çetinkaya-Rundel, M. (2021). Web scraping in the statistics and data science curriculum: Challenges and opportunities. Journal of Statistics and Data Science Education, 29(sup1), S112-S122.
[11] E. Suganya, S. V. (n.d.). Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms. International Conference on Intelligent Systems Design and Applications. 2019.
[12] Farley, E.J. and Pierotte, L. "An Emerging Data Collection Method for Criminal Justice Researchers." Justice Research and Statistics Association, December 2017
[13] Fiesler, C., Beard, N., Keegan, B. C. (2020). No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. Proceedings of the International AAAI Conference on Web and Social Media, 14(1), 187–196.
[14] Georgescu, T. M., Smeureanu, I. (2017). Using Ontologies in Cybersecurity Field. Informatica Economica, 21(3).
[15] Gheorghe, M., Mihai, F.-C., Dârdală, M. (2018). Modern techniques of web scraping for data scientists. International Journal of User-System Interaction, 11, 63–75.
[16] Grasso, G., Furche, T., Schallhart, C. (2013). Effective Web Scraping with OXPath. Proceedings of the 22nd International Conference on World Wide Web (pp. 23–26). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2487788.2487796
[17] Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., Perez-Meana, H. (2018). A web scraping methodology for bypassing twitter API restrictions. arXiv preprint arXiv:1803.09875.
[18] Hillen, J. (2019, November). Web scraping for food price research. British Food Journal, 121, 3350–3361.
[19] Kasereka, H. (2020). Importance of web scraping in e-commerce and e-marketing. SSRN Electronic Journal.
[20] Krotov, V., and Tennyson, M. 2018. "Scraping Financial Data from the Web Using the R Language," Journal of Emerging Technologies in Accounting, Forthcoming.
[21] Krotov, V., Silva, L., 2018. Legality and ethics of web scraping, Twenty-fourth Americas Conference on Information Systems, New Orleans..
[22] Landers, R. N., Brusso, R. C., Cavanaugh, K. J., Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological methods, 21(4), 475.