Web Scraping

VOOTLA GNANESWAR, VURIBINDI VENKATA HARI PRASAD, YENIGALLA CHAITANYA SAI

*Department of CSE*

[*vootlagnaneswar@gmail.com*](mailto:vootlagnaneswar@gmail.com)*,*

[chaitanyasaiyenigalla@gmail.com](mailto:chaitanyasaiyenigalla@gmail.com),

[harivuribindi@gmail.com](mailto:harivuribindi@gmail.com),

***Abstract—this article is about web scraping which is used to scrape information from the website and store the information in excel sheets.******The website spiders to be scraped is operated upon by spiders which we create using AI which function as bots. The website spiders to be scraped is operated upon by spiders which we create using AI which function as bots. A web crawler is a piece of code that travels the Internet and collects data from various web pages, also known as web scraping. Web scraping replaces the need for manual data entry and more easily reveals trends among data collected. Main objective of web scraping is to extract information from one or many websites. Web crawlers then process the information into simple structures such as spreadsheets, database.***

***Keywords*** *—* ***web scraping, web crawler, spider, artificial intelligence.***

# I.INTRODUCTION

The World Wide Web consists of an interlinked network of information, which is presented through websites to the users. World Wide Web has significantly changed the way we share, collect, and publish data. The amount of presented information grows constantly. Also with the usage of Web as a new marketing and sales channel the quantity of content multiplied. Online merchants offer large packs of data to describe their products.

Knowledge base providers offer access to their databases. With this unorganized growth, it is no longer possible to manually track and record all available sources. That moment, is when Web Scraping evolved. Automated techniques allow the collection of a massive amount of data from the Web compared to manual data extraction. Together with Web Scraping another term became very important – Meta Data. Massive collection of data obtained by Web Scraping allows Meta Data analysis.

As data grows in amount, variety, and importance, business leaders must focus their attention on the data that matters the most. Not all data is equally important to businesses or consumers. The enterprises that thrive during this data transformation will be those that can identify and take advantage of the critical subset of data that will drive meaningful positive impact for user experience, solving complex problems, and creating new economies of scale. Business leaders should focus on identifying and servicing that unique, critical slice of data to realize the vast potential it holds. ((IDC), 2017) David Reinsel John Gantz John Rydning, Data Age.

# II.LITERATURE REVIEW

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.

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.

III PURPOSE OF WEB SCRAPING

Enormous amounts of source information, available on the World Wide Web, are still in the format of a Hypertext Markup Language (HTML) page. Automated extraction is difficult, because the intended reader was a human. This chapter introduces the motivation and purpose of Information extraction through Web Scraping. These utilizations are often only possible because the existence of automated Web Scraping. Without these techniques, it would be impossible to collect the amount of data repeatedly and in reasonable time.

Rapid growth of the World Wide Web has significantly changed the way we share, collect, and publish data. Vast amount of information is being stored online, both in structured and unstructured forms. Regarding certain questions or research topics, this has resulted in a new problem—no longer is the concern of data scarcity and inaccessibility but, rather, one of overcoming the tangled masses of online data. (B.C., 2016)

IV Market analysis and research

Data collection from online sources became one of the market research methods. It offers much faster response, compared to a classical surveying. While people considers it best to utilize traditional surveys, Web-scraping is seen as cost effective support for such instruments. To get comprehensive picture and to gain knowledge of the tools in markets multiple sources should be used.(Raulamo-Jurvanen P., 2016)

Consumers are active in the online world and share their experience, frustration or motivation. Companies that wish to learn more from consumers can add online sources of information. Web scraping is one of the method to collect such data. Targeted data collection from e-shop and advertising servers helps to update Indexes. Which are based on frequently changed prices. Indexes built through automated Web scraping can offer more frequent update intervals.

With the increasing relevance and availability of on-line prices that we see today, it is natural to ask whether the prediction of the consumer price index (CPI), or related statistics, may usefully be computed more frequently than existing monthly schedules allow for Wegmann and Chapple (2013) used a small sample of 338 Craigslist listings to study the prevalence of secondary dwelling units in the San Francisco Bay Area. Finally, Feng (2014) web-scraped 6,000 Craigslist listings to study Seattle’s housing market. (Daniel Glez-Pen, nedatováno)

V Methods of Web Scraping

The methods of Web Scraping evolved together with the World Wide Web. Not all listed methods were available at the beginning. There are two examples to mention, because these are presently the most used techniques.

Since 2000 the Document Object Model (DOM) became more popular in DHTML. A broader acceptance later on allowed the HTML Parsing technique to evolve to DOM Parsing.

Second example are Application Programming Interfaces (APIs). This technique is the youngest on the list, the growth of available content APIs is dated from 2005. According to ProgrammableWeb.com the number of APIs has grown within 8 years from 0 to 10302. (Berlind, 2015)

VI ANALYSIS OF WEB SCRAPING

Web scraping can be performed by teams in the following steps shown in figure 2. The scraping organisations take the websites details from the clients from where the data to be extracted and analysis in done by the experts. Then they get it approved by the clients. After approval the extraction process is done for the required data along with data configuration and then the final information is delivered to the client followed by collecting the feedback.

Diagram

Description automatically generated

Figure 2: Phases of Web scraping

VII **ANALYSIS AND DESIGN**

**SPIDERS**

Spiders are classes which define how a certain site (or a group of sites) will be scraped, including how to perform the crawl (i.e. follow links) and how to extract structured data from their pages (i.e. scraping items). In other words, Spiders are the place where you define the custom behavior for crawling and parsing pages for a particular site (or, in some cases, a group of sites).

For spiders, the scraping cycle goes through something like this:

* You start by generating the initial Requests to crawl the first URLs, and specify a callback function to be called with the response downloaded from those requests.
* The first requests to perform are obtained by calling the **start\_requests()** method which (by default) generates **Request** for the URLs specified in the **start\_urls** and the **parse** method as callback function for the Requests.
* In the callback function, you parse the response (web page) and return item objects, **Request** objects, or an iterable of these objects. Those Requests will also contain a callback (maybe the same) and will then be downloaded by Scrapy and then their response handled by the specified callback.
* In callback functions, you parse the page contents, typically using Selectors (but you can also use BeautifulSoup, lxml or whatever mechanism you prefer) and generate items with the parsed data.
* Finally, the items returned from the spider will be typically persisted to a database (in some Item Pipeline) or written to a file using Feed exports.

Even though this cycle applies (more or less) to any kind of spider, there are different kinds of default spiders bundled into Scrapy for different purposes. We will talk about those types here.

VII TECHNIQUES

Techniques Web scraping is the method of dynamically extracting data or gathering World Wide Web information. This is a sector with effective advances having a shared aim with the object model mission, an ambitious project which still needs innovations in word processing, conceptual comprehension, artificial intelligence and communication between humans and computers.

• Manual copy and pasting This is simple copy pasting method that supports in some situations very efficiently.

• Pattern matching This is the proficient approach to extract information from websites.

• Web mining solutions This is one of the commercial method utilized for extracting data from various websites.

• HTTP programming The method used for retrieving data from dynamic and static web pages.

• Parsing through HTML Parsing HTML web pages to recover information and renovate it.

• Web data extractor Data can be scraped from websites by only some clicks

• Web scraper This is an easy and free tool and a chrome extension used for extraction of data.

• Mozenda This is a web-scraping application focused on enterprise clouds. It has an application of point-to-click and a customisable UI.

• Scrapy This is free software of web scraping application used to create web crawlers in Python. It gives all of the resources necessary to easily collect, process, and store data from websites in desired format and structure.

• Data scrapper This is an easy web scrapping tool for information extraction. This is an extension to a specific website that lets to convert information into a clear tabular format

# VIII.ACKNOWLEDGMENT

This study is based on the author's mini project from his degree programme in 2022, which he completed with the help of Xeraphin (Asst. Prof.) from Bharath University in Chennai. The author of this document does not claim ownership of any algorithm, code, data, formulas, definitions, or problemsolving method. He has simply used his head to put it all together, and he has worked hard to collect findings and put it all together in the format of an IEEE paper.

IX**. 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 Banking Sector Online media Consultancy managemet Insurance Network and security Marketting Finance Internet Computer software IT Sector 0 10 20 30 40 50 60 Small websites Average websites Large websites .

X**.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] [www.scrapy.org](http://www.scrapy.org)