JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EFFICIENT SCRAPING OF DATA FROM WEBSITES USING SELENIUM

[1]Shreya V. Dhoke, [2]Anupama D. Sakhare,[3]Satish J. Sharma

[1]Student, [2]Assistant Professor, [3]Professor

[1,2,3]Department of Electronics and Computer Science

[1,2,3]Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur, India

Abstract: Internet is an ocean of information spread across various websites, where it is categorized, interlinked and mostly freely available for everyone. A vast amount of data is being created every second. All this 'Big Data' is in heterogeneous formats. We need to access information fast and quickly. Data extraction can be done manually but it can be time-consuming and can also be a very complicated task, for this reason Web Scraping is used. Web Scraping is the technique of automating the process of navigating through links, and then navigating and collecting the relevant data from these relevant links. The proposed system is a method of extracting and restructuring information from web pages. It is a technique for targeted, automated extraction of information from websites. This system acquires non-tabular or poorly structured data from websites and converts it into a usable structured format. The main objective of the proposed system is to extract information from one or many websites and process it into simple structures such as CSV files. In this proposed system, Text Grepping technique is used, that offers insight into price data, market dynamics, prevailing trends, practices employed by various competitors, and the challenges they face. The result of this technique is to easily access relevant data from websites. The proposed system can be modified for scraping dynamic websites. This proposed system will be beneficial in many business and at education areas.

Keywords: Web scraping, Big Data, CSV file, Structured and Unstructured data

I. INTRODUCTION

Internet contain various information from various websites, where it is categorized, interlinked and freely available for everyone. Some data that is available on the web is presented in a format that makes it easier to collect and use it. For extracting relevant data from websites, it is very tedious task and time consuming to manually extracting it. For formatting this Web Scraping is used. Web Scraping is process of extracting relevant data from websites. This technique of automating the process of navigating through links, and then navigating and collecting the data from relevant websites. After automation, instead of manually coping the data from websites, Web Scraping will replicate the same task within a fraction of time. The various technique used for Web Scraping are Text pattern matching, HTTP programming, DOM parsing, Text Grepping, Vertical aggregation, Semantic Annotation recognizing, computer vision web-page analysis etc. Most of required data is unstructured data in HTML format which is then converted into structures data in a spreadsheet or a database so that it can be used in various applications. Internet contain various information from various websites, where it is categorized, interlinked and freely available for everyone. Some data that is available on the web is presented in a format that makes it easier to collect and use it. For extracting relevant data from websites, it is very tedious task and time consuming to manually copy-paste it. For this Web Scraping is used. Web Scraping is process of extracting relevant data from websites. This technique of automating the process of navigating through links, and then navigating and collecting the data from relevant websites. After automation, instead of manually coping the data from websites, Web Scraping will replicate the same task within a fraction of time. The various technique used for Web Scraping are Text pattern matching, HTTP programming, DOM parsing, Text Grepping, Vertical aggregation, Semantic Annotation

recognizing, computer vision web-page analysis etc. Most of required data is unstructured data in HTML format which is then converted into structures data in a spreadsheet or a database so that it can be used in various applications.

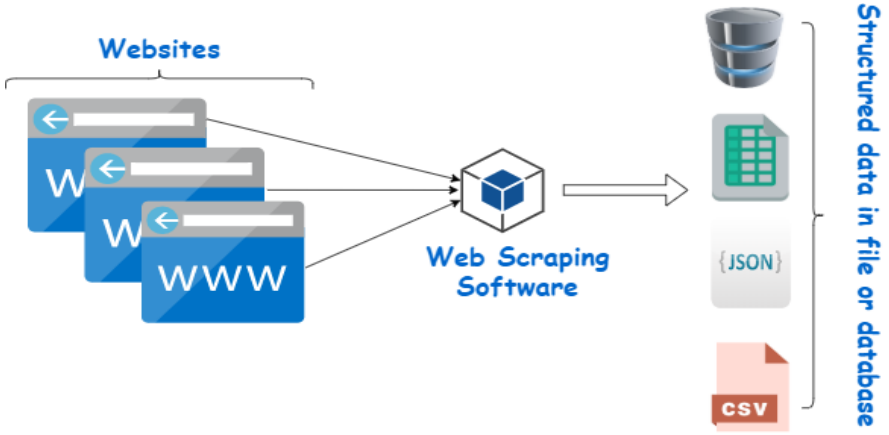

Fig 1: Web Scraping Structure

II. OBJECTIVES

The objectives to be achieved in this project are:

- To acquire non-tabular or poorly structured data
- To scrap data from other sites
- To verify the possibility to produce statistical outputs using predicted data.

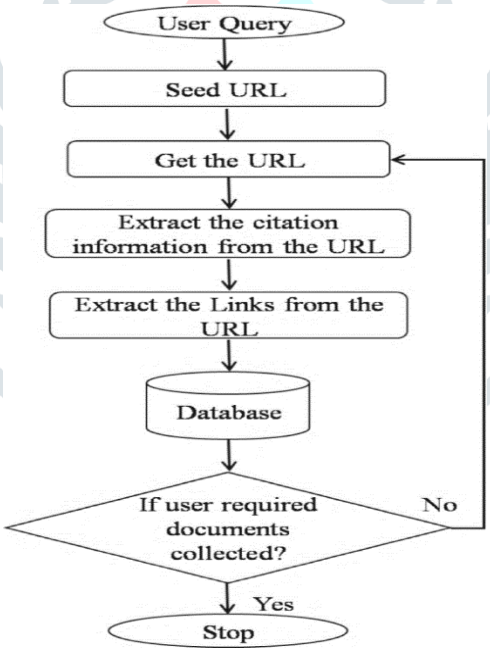

Fig 3: Complete Flowchart of Web Scraping

III. REVIEW OF LITERATURE

Web Scraping, i.e. the automated and targeted extraction of data, is a traditional technique to retrieve Web content at scale. A multitude of frameworks and Application Programming Interfaces to develop customized scrapers, as well as configurable ready-to-use scraping tools exist.

Renita Crystal Pereira [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 [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. 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.

Sameer Padghan [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[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 [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 [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 methodThe 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.

To know how the data extraction process has evolved has so much one must understand the techniques involved in this method of web scraping is important scraping has been around nearly as long as the web. The impact behind business web scraping has dependably been to pick up a simple business advantage and incorporate things like undermining a contender's special valuing, taking leads, commandeering promoting efforts, diverting APIs, and the inside and out robbery of information.

III. EXPERIMENTAL WORK

The proposed system can perform the automation tasks according the given points:

1. The web is filled with text. Most text, though, is structured according to HTML or XHTML markup tags which instruct browsers how to display it. These tags are designed to help text appear in readable ways on the web and like web browsers, web scraping tools can interpret these tags and follow instructions on how to collect the text they contain.

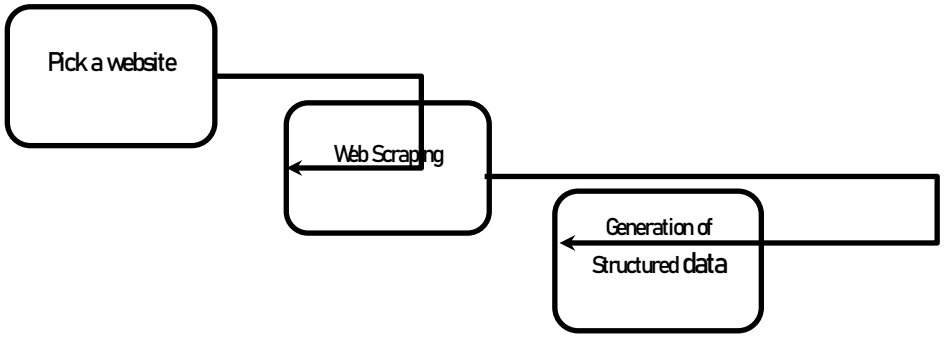

Fig 2 : Web Scraping process

2. Web scraping tools can range from manual browser plug-ins, to desktop applications, to purpose-built libraries within Python language.
3. A web scraping tool is an Application Programming Interface (API) in that it helps the client (you the user) interact with data stored on a server (the text).
4. Selenium library is used to connect with Web drivers to browser plug-in tools of chrome or Firefox browser.
5. Selenium automation tool stimulate the automation on relevant link and generate it into CSV file.
6. Unstructured data can be obtained in structured format in CSV file or a spreadsheet.
7. After generation of CSV file one can analyze the data accordingly.

IV. CONCLUSION

The proposed system reduces the manual work of extracting of data in less amount of time. Automatically generation of data in required file makes work easy to analyze it.

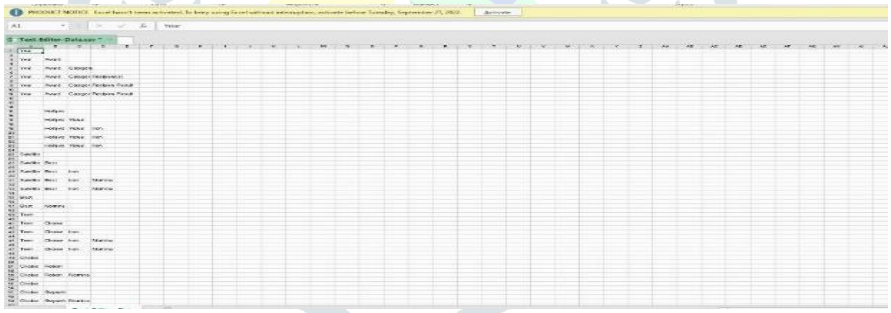

Fig 3: CSV file of Desired Output

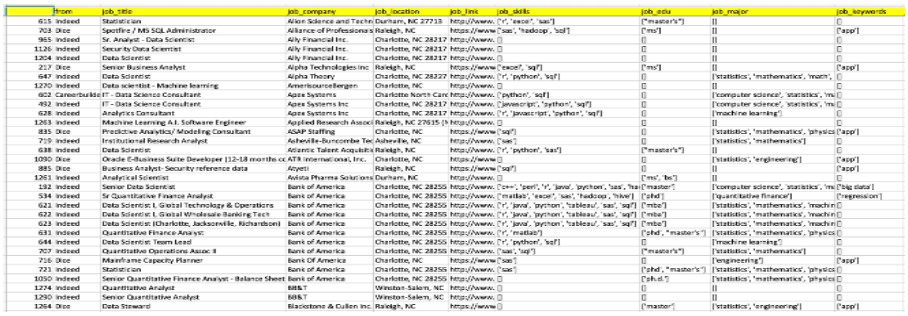

Fig 4. CSV file

V. REFERENCES

[1] Renita Crystal Pereira and Vanitha T, *"Web Scraping of Social Networks,"* Int'l J. of Inno. Res. in Comp. and Comm. Engg., 3(1), 237-240, 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] David Mathew Thomas, Sandeep Mathur ,Amity Institute of Information Technology ,Amity University (AUUP), Sec-125, Noida

[10] http://wthtjsjs.cn/gallery/1-whjj-june-541,pdf case study.

[12] 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.