### 1. INTRODUCTION

### 1.1 Brief Information about the project:

### The web is a major source of information for many professionals in various sectors. It contains useful and useless, structured and non-structured information, in different formats, and from various sources. However, in addition to being a very complex activity, Web Scraping is a time- and resource- consuming task, especially when it is carried out manually. This complexity increase depending on data and collection websites. Many techniques have been used to retrieve content from a web page: Cut/Paste, http, Query languages for semi-structured Data, DOM or even Web-Scraping. Many advanced techniques are also used to collect data from the web. Among these, one can mention API computer languages, robots, intelligent agents and Web Scraping. A web scraper is, therefore, a software that simulates human browsing on the web to collect detailed information data from different websites. The advantage of a scraper resides on its speed and its capacity to be automated and/or programmed.

### However, no matter what technique is used, the approach and the objectives remain the same: capture web data and present it in a more structured format. This paper revisits existing frameworks for use, approaches, categories, and tools of Web Scraping through the identification of their strength and limits.

### 1.2 Motivation of project:

### Web Scraping is a process of automating the extraction of data in an efficient and fast way. With the help of web scraping, you can extract data from any website, no matter how large is the data, on your computer. Moreover, websites may have data that you cannot copy and paste. When you extract web data with the help of a web scraping too, you would be able to save the data in a format such as csv. You would then be able to retrieve, analyze and use the way you want. So web scraping simplifies the process of extracting data, speeds it up by automating it and creates easy access to the scrapped data by providing it in a CSV format.

### 1.3 Objective of the project:

Web scraping is a technique for targeted, automated extraction of information from websites. Similar extraction can be done manually but it is usually faster, more efficient and less error – prone to automate the task. Web scraping allows you to acquire non-tabular or poorly structured data from websites and convert it into a usable, structured format, such as a .csv file or spreadsheet. We are also able to scrape job requirement data from LinkedIn. LinkedIn is very specific about what gets written to the screen at any given time. For instance, new jobs only load after the user has scrolled to the bottom of the page so that the complete data gets loaded, we will scrape the data we want.

### 1.4 Organization of the project:

* **Chapter2: Literature Survey:** This chapter mainly consists of background of the project, possible approaches, introduction and comparison of technologies.
  + - **Chapter3: System Analysis:** This chapter mainly consists of description of current system, proposed system, and requirement specifications.
    - **Chapter4: System Design:** This chapter mainly consists of module descriptions and algorithms with examples and unified modeling language diagrams: use case diagrams, class diagrams, sequence diagrams, collaboration diagrams, and activity diagrams.
    - **Chapter 5: Technology Description:** This chapter mainly consists of technology used in this project.
    - **Chapter 6: Sample code:** This chapter mainly consists of sample code for few modules.
    - **Chapter 7: Testing:** This chapter mainly consists of testing techniques and test cases for modules
    - **Chapter 8:Screenshots:**This chapter mainly consists of output screens of this project
    - **Conclusion:** Main conclusion of this project.

# **CHAPTER-2**

# **LITERATURE SURVEY**

## **2. LITERATURE SURVEY**

**Kaushal Crystal Pereira ET.** al., 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., 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., 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., 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 web- based 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., 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., 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** 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., 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 liberty- related 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.