IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey on Web Scraping and its Applications

Prof.Shivsagar Gondil
Computer Engineering
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India
shivsagar143@gmail.com

SmitPatne
B.E., Computer Engineering
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India
smithp17999@gmail.com

Tejas Raut
B.E., Computer Engineering
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India
tejaskraut1999@gmail.com

Vinit Bhagat
B.E., Computer Engineering
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India
vinitbhagat02@gmail.com

***Abstract*-In the manufacturing world, competitive research using site scraping would be an important method for consumers to locate the desired commodity at the best price. It will be solely focused on a comparison of various online sites (with preference given to none) for a single smartphone product chosen by the customer. It will also recommend the most commonly purchased or trending mobiles within the user-specified price range. This would raise consumer income while still saving time. Comparator Analysis using web scraping will be an effective tool in the industrial world which will help the users to find the desired product at best price . It will totally be based on comparison between multiple online platforms available (with favoured to none) for a particular mobile product selected by the user. Furthermore it will also suggest most frequently bought or the most trending mobile at the user selected price range . This will result in customer profits and will also be time efficient.**

Keywords— Web Scrapping and Crawling, Price Comparison, Big Data, Business Intelligence, Web Extracting.

I. INTRODUCTION

One of the main problem that the users face while e-shopping is the problem of plenty. There are multiple online platforms available for a particular product but the problem is which one services the best product at most reasonable price. Comparator Analysis tool comprises simple web scraping techiniques and django(python). All the websites that are being scraped are completely legal. Multiple forms of data is stored in different different forms. The internet world is actually immense in terms of the number of web pages and a vast number of informational resources accessible in different formats such as print, multimedia, audio video, etc., which will be based on the inconsistency of repossession due to the insignificance of the fact that the customer is observing the details shown on websites can only be correct. A web browser was used to access this page. They don't have the requisite features. to make a personal copy of this knowledge Then the only choice is to is to manually copy and paste the data from the website into a text document, browser into our computer's hard drive, which is very huge.

The method of interrogating and analysing data in order to identify answers to challenges is known as data processing. Discovering problems, resolving the availability of suitable data, determining which method can assist in finding a solution to the interesting problem, and communicating the result are all part of the analysis process. The data must be separated into various steps for analysis, such as starting with its specification, assembling, organising, cleaning, re-analyzing, applying models and algorithms, and finally arriving at a final result.

II. APPROACHES, CATEGORIES AND TOOLS

A. Mimicry Approach

This form of scraper works by using pre-programmed custom rules. The scraper is preconfigured with the location of the data to be obtained from a web page. This mechanism is used for DOM selectors originating from click-based leaning. Because of its neatness, this technique is reasonably ,effective, but it is less suited to processing multiple heterogeneous websites. In addition, if the source website's graphic design changes, the engine should be reprogrammed to locate the necessary details. This technique is used by tools like Import.io and Mozenda.

B. Weight Measurement approach

This procedure is based on a generalised algorithm that evaluates a web page's DOM tree and determines the weight of words in each branch. The algorithm selects the node as the starting point for the main text and removes the text from all of the child nodes using deduction. The key benefit of this process is that it does not require much testing and

will respond to changes in the source websites' graphic design. The findings, on the other hand, are usually very loud.

C. Differential approach

This technique is based on the fact that the content of two pages from the same website can only vary in the page's body.
The menu bars, right or left columns, and footers are expected to be perfectly similar for two pages on the same website, according to this logic. Previously, the process consisted of using a masking algorithm to superimpose the two pages by only eliminating the variations.

D. Machine learning approach

The key idea of a wide sampling of Web pages is to train an algorithm. Machine learning is based on geographic indicators of text blocks on the page: statistical measures are carried out in relation to each other's text blocks where the main text block is located. The computer will then deduce where the text is generally placed by itself. The more precise the algorithm is, the greater the sample.

CATEGORIES AND TOOLS

a) Ready-made Tools

i. Browser extensions

- Spider: It is a Google Chrome extension that is available for free. Each column on the screen represents a different type of element that can be retrieved. To add an item to a column, all you have to do is click on it. The output can be downloaded in JSON or CSV format.
- Data Scraper: It is a Google Chrome extension that extracts data from web pages and exports it in CSV and/or XLS formats.
- Agenty: It is a Chrome extension that allows you to quickly and easily extract data from a web page using CSS classes.
- Data Miner: It is a Google Chrome extension that extracts data from web pages and saves it as a CSV file or an Excel spreadsheet. For more than 15,000 websites, there are more than 50,000 free predefined queries.
- Cloump U-Scrape: It's a Firefox extension. With the help of a free Excel add-on, it converts web pages into API and connects them via Excel. It can process complex nested data structures and supports the JSON schema.
- OutWit Hub.It is a robust Firefox add-on. It makes it simple to extract links, images, email addresses, data tables, and other items. Images and documents are saved directly to your hard drive, while data can be exported to databases, CSV, HTML, or Excel files.
- Dexi.io It is a website that allows you to share your ideas with others. This Firefox add-on allows you to programme robots and retrieve data in real time. It can collect information from any website.

ii. Software and Platforms

- Import.io: a table-based function library enables the end user to create tailored formulas to enrich all the data.
- Web Extract Easy: It is a scraper with multi-format results written using.NET technology (Excel CSV, text, XML, HTML ...). The time it takes to extract is one of the limitations of this tool.
- Web Info Extract: A scraper capable of storing information in a database. When set up, the scraper monitors the website constantly and, depending on the changes, the task assigned to the tool is updated when new content is added to the page.
- Fminer: It's one of Python's most successful visual web scraping tools. The scraping flux and actions have a nice schematic representation. It also enables custom python code to be executed.
- Weboob: The application includes QBooblyrics, QBooobMsg, QCineoob, QCookboob, QFlatBoob, QHandjoob, Q HaveDate, QVideoob and QWebConfig-Edit all written in python.
- PySpider: a Python-based web robot. It supports JavaScript pages and has an architecture distributed. One of PySpider's advantages is its easy-to-use interface. It is possible to store the data in JSON and CSV formats.

b) Libraries of programming Language

The libraries of programming languages PHP, Java, Python, and NodJs make up the second group of web scraping software. Python puts together Lovely Broth, Newspaper, Lxml, and other materials. Jsoup, Jaunt, StormCrawler, Norconex http, Collector, and other Java API web scraping tools are available. Cheerio and Apify are two NodJs solutions. In addition to these simple libraries, there are various implementation examples (Framework) that can be used to easily build a scraper; that is, the scraper's remaining function.
These solutions are for professionals collecting information or in Automated Natural Language Processing in all situations (NLP). The first thing to remember about these methods is that the vast majority of them are not intended for laypeople in that area and are not domain-specific. As a consequence, configuring these libraries for a given domain often necessitates the use of an overlayer. Newspapers are designed to extract news stories in a foreign language, such as English.

III. EXISTING SYSTEM

The following is a list of fields that use Web Scraping:

A. Retail and Manufacturing

1. *Price comparisons with rivals*- Price is highly significant in today's e-commerce world. Keep an eye on the rivals' pricing tactics. Trying to keep track of prices by hand, on the other hand, is not a feasible choice. Furthermore, costs fluctuate from time to time. As a consequence, manually keeping track of prices becomes almost impossible. This is where web scraping joins the image. It automates the method of collecting competitor rates and keeps you posted about new pricing tactics used by your rivals. It can be done in no time and as many times as you like with web scraping.

2. *MAP Enforcement is being tracked.*- Manufacturers will have to keep an eye on retailers to see if they are adhering to the minimum offer. It is not feasible, however, to visit each website and search for the same. In this situation, web crawling comes to their assistance. Manufacturers can quickly and efficiently track MAP enforcement using site scraping. They can do so without losing a lot of time because web scraping can create this information in a matter of seconds.

B. Equity and Finance Research

1. *Aggregated News Articles*- News is a significant source of knowledge in the areas of banking and insurance. It is, however, difficult to read every newspaper and article by hand. As a result, site crawling is used to derive useful inputs from numerous news stories, headlines, and other outlets in order to turn them into actionable investment insights.

2. *Aggregation of Business Statistics*- Although there is a wealth of consumer data on the Internet, it is spread across tens of thousands of websites. You can browse and scan the search results, but it is inefficient and time-consuming. In terms of equity analysis, database scraping is used to crawl data from numerous websites and derive actionable intelligence from them.

C. Data Science

1. *Analytical Prediction*- Predictive research is the method of reviewing current data to discover patterns and model potential events or developments. Predictive forecasting cannot predict the future with accuracy, but it can forecast the odds. Predictive analysis has applications in a number of areas, including industry. Predictive research is used to investigate and comprehend consumer actions, goods, and a number of other variables in order to assess risks and opportunities. However, as is obvious, it is a form of study that is carried out using a huge volume of existing data.

2. *Natural Language Processing*- Natural language processing, in comparison to programming languages such as Python, is the method of allowing computers to comprehend the natural languages spoken by humans. Natural language processing is well-known for its use in emotion analysis. Data scientists use social media feedback to process and determine how a brand is doing. For any NLP programme to operate, computers would require access to vast volumes of data. Web scraping is one of the few useful approaches for scraping and translating data from social media comments or other sources into an accessible format. As a result, site crawling has become increasingly important as NLP has increased in popularity.

D. Risk Management

When it comes to hiring new staff or working with new customers, there are some risks to consider. Without any risk management techniques, it is difficult to disregard the risk.As a result, companies regularly perform background checks on prospective hires, employers, and customers. However, since it requires reviewing many different sources of records, such as press and news papers, sanctions lists, corporate registers, legal files, barred directors lists, insolvency registers, financial registers, and so on, it may be a time-consuming and cumbersome activity. Site scraping techniques are used to retrieve data from the above sources easily and process it so that the background checks can be done.

E. Product, Marketing and Sales

1. *Centred on evidence promoting*- For any marketing or distribution endeavour, data is important. It isn't a brand-new idea. Access to info, on the other hand, may often differentiate two advertisers. Web scraping will include the details required to devise strategies.

2. *Marketing of Content*- Web crawling is used in content marketing to gather data from numerous websites such as Twitter, Tech Crunch, and others. This detail can then be used to generate interesting content. As you might be aware, engaging content is important for business development and web traffic.

3. *Study of Rivals*- When it comes to comparative research, it can be difficult to collect all of the details you need from multiple sources in order to compile a thorough comparison and understanding of your rivals. It is not just time intensive, but it is also unreliable since you can or might not be able to access data correctly. In today's fast-paced dynamic environment, speed and precision are important. Web scraping, on the other hand, has changed this area and created a more effective solution by easily extracting data and encouraging comparative research. Web crawling is used in this manner to simplify data retrieval and competitive analysis.

F. Other Industries

1. Academics

The academic world is highly dependent on data. Academic work is mainly focused on one kind of data or the other. Academics must collect data and then process it in order to arrive at the requisite insights, whether it be for a teaching assignment or a research project. They can now retrieve and process the data they need even more effectively thanks to web scraping.

2. Non-Profit Corporation

Non-profit organisations, too, need data in order to identify their mission and promote their work. Site scraping apps make it easy for them to extract the details they need to find out their priorities and results, enabling them to press on with their noble ventures.

3. Search Engines

Web Scraping is a core component of career search engines like Teambuilder. Web crawling is also used by many other collective suggestion programmes. Grundy, GroupLens, Video Recommender, and Ringo are examples of applications that used collective filtering algorithms to simplify recommendations first.

The table below shows the target and solution that were reached in the published articles that the authors used as a reference.

Target	Solution
To assemble computer parts at best price and quality.	Formulation of a web based application that will help the user to get the recommendation and at best price . This application compares data of five local shops and is based on Pentaho software , java script and SQL as data base.
Ingredient / recipe algorithm using web scraping fro smart chef	A python based web scraping helps to collect requirements automatically customer can maintain health according to the diet plans and enjoy variety of dishes. Scrapped website – **www.foodnetwork.com .**
Climate analysis using web scraping	A research by Charmaine Bonifaciio etal (2014) introduced to Canadian Climate Data scraping Tool where data was obtained directly from weather forecast organization of Canada

IV. FUTURE SCOPE

The problems that lie ahead include the web's nonuniform structure, which is a complex space of irregular knowledge institutions and structure. Owing to this lack of accuracy, collecting information in a machine-meaningful arrangement can be a challenging theory problem that gets worse with size when you need structured knowledge method a.k.a. information extraction this spot test when a large number of subtleties are needed to infiltrate to a specific plan from a large number of sources this can be overwhelmed by a large number of subtleties are required to infiltrate to a particular plan from a large number of sources this can be overwhelmed by a large improvement in health and condition component arrangement. Indeed, even with all of the confinement's site knowledge, there are always possibilities for use if we know how to bring it to the best possible use.

V. CONCLUSION

Web Scraping is useful for analysis of various online platforms for a product which is beneficiary for the customer. Major components of this tool are it provides a single platform , lot more choice and cost efficient system site scraping assists in the automated processing of data in compliance with the user algorithm's specifications. As the Internet has expanded rapidly and companies have become highly focused on results, having access to the most up-to-date information on any particular topic has become a necessity. data has become the backbone of all decision-making processes. As a result, web scraping has found use in nearly every significant project in recent times. It's apparent that web scraping technique in an innovative and sophisticated manner will be able to get ahead of the market and achieve a strategic edge.

VI. REFERENCES

[1] A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages ERDINÇ UZUN
Department of Computer Engineering, Çorlu Faculty of Engineering, Tekirda§ Nam_k Kemal University, 59860 Tekirda§, Turkey

[2] Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef
Shilpa Chaudhari, Aparna R., Vinay G Tekkur, Pavan G L., and Shreekanth R Karki
Department of Computer Science and Engineering, S. Ramaiah Institute of technology, Bangalore-560054

[3] THE USE OF WEB SCRAPING IN COMPUTER PARTS AND ASSEMBLY PRICE COMPARISON
Leo Rizky Julian1, Friska Natalia2, Department of Information System, Universitas Multimedia Nusantara, Boulevard Street, Gading Serpong, Tangerang, Banten 15810, Indonesia

[4] Web Scraping: State-of-the-Art and Areas of Application, Rabiyatou DIOUF, Edouard Ngor SARR, Ousmane SALL, Babiga BIRREGAH, Mamadou BOUSSO, Sény Ndiaye MBAYE, Université de Thies, Thies, SENEGAL, Université de Technologie de Troyes, Troyes, France, UCAO-Saint, Michel, Dakar-SENEGAL

[5] Data Analysis by Web Scraping using Python, David Mathew Thomas, Sandeep Mathur
Amity Institute of Information Technology, Amity University (AUUP), Sec-125, Noida

[6] A Review on Web Scrapping and its Applications, Vidhi Singrodia, Anirban Mitra, Amity University, Newtown, Kolkata, Subrata Paul, Research Scholar, MAKAUT,Kalyani, NADIA