Comparative Study Of Various Scraping Tools: Pros And Cons

Priya Matta
Department of Computer Science and Engineering
Graphic Era Deemed to be University, Dehradun, India
mattapriya21@gmail.com

Sonal Sharma
Department of Computer Application
Uttaranchal University,
Dehradun, India
sonal_horizon@rediffmail.com

Nitin Uniyal
Department of Mathematics
University of Petroleum and Energy Studies,
Dehradun, India
nuniyal@ddn.upes.ac.in

Abstract. ***As technology keeps on improving and data is still considered the topmost priority, people need to know how they can make the use of old data available, and impact their worlds. The data these days may be raw data, refined data, structured or even the unstructured data. These data finally result into a huge amount of data with varying volume and velocity, generating the concept of Bigdata. Data Extraction has shifted the way we view the world and as it is so necessary, we need to know certain tools available for this service. Web scraping has become so important for businesses to grow, scientific research, or even to get the knowledge of the most read article on the internet. In this paper, we cover the most powerful known web scraping tools available to date and study them. We have compared these tools and clarified their benefits as well as drawbacks concerning various applications.***

Keywords: structured data, bigdata, web scraping, data extraction, web scraping tools

I. Introduction

Over the past few decades, web scraping a huge positive impact on businesses, marketing, scientific fields, and even our daily lives. Data always plays an active role in making all the fields flourish and is also needs to be analyzed to maintain its demand. For that, either people will tend to copy it from the websites which are forbidden mainly is all the cases, otherwise, the data copying can itself be a tedious task which will also increase with the case of dynamic websites. Due to this scenario web scraping was introduced, web scraping, in general, is a technique that helps to extract useful data from websites manually (by a program) or by tools(software) that is later analyzed depending on the needs of the user. Many new emerging businesses can succeed because of web scraping as lead generation has become a simple task. Web scraping is a powerful technique that anyone can make full use of its resources. In this paper, we study various tools and techniques that are available to date to extract data, also keeping in mind all their advantages and disadvantages and how they suit the best. Later we come up with an algorithm that will serve as an upgrade for the process of web scraping.

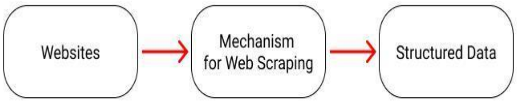
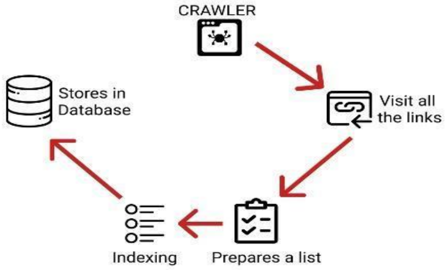


Fig. 1. Basic web scraping architecture

When it comes to data extraction, in general, a lot of techniques stay at crossroads, like Data Mining, Information Extraction, and Information Retrieval. Information extraction is the extraction of structured data from an unstructured, semi-structured readable format. Extraction is possible from either database, documents, and even websites. Whereas information retrieval offers searching techniques on multimedia resources. Both are mainly used for the extraction of unstructured data and later used for data mining purposes. Web scraping offers extraction of data from web sites. We can do web scraping either manually by writing the code for the scraper or by using a tool that will do the same without coding. Tools offer a wide range of web scraping applications beyond just extracting text, also providing the files in CSV file or any format possible.

Along with Web scraping, a common term comes on the surface, i.e. web crawling. A web crawler is termed as an internet bot that searches WWW for web indexing. According to Shrivastava [1], the task of a web crawler is, "to maintain and manages the index the web pages and make searching fast, truthful, and productive. Pages visited by the web crawlers are copied in web repository for later use. The crawler is a tool to collect and keep the database up-to-date."



2022 IEEE Delhi Section Conference (DELCON) | 978-1-6654-5883-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/DELCON54057.2022.9753358

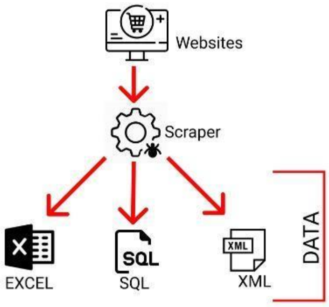


Fig. 2. Difference between Web Scraping and Web crawling.

II. Evolution Of Web Scraping

The world wide web came into existence in 1989, and the first web robot, “worldwide web wanderer,” was built in June 1993, which remained only to estimate the web’ size. It was in the year 1993, December, when the foremost crawler-based web search engine, JumpStation, was launched. During that time, the number of websites on the web was less in number. Hence the search engines relied on their website administrators to manage and update the links/URLs in an appropriate format. In contrast, JumpStation was the only search engine that made a new jump relying on a web robot. In the year 2000, the first Web API and API (Application Programming Interface) crawler came, which made things much more comfortable to develop a program by rendering the building blocks. In the very same year, Salesforce, as well as eBay, began their API, which enabled the programmers to access and download the public data. According to Wikipedia [2], with the elapse of time, several websites happen to provide web APIs for users to access their public database.

Before the birth of Yahoo! in 1994, many search engines were created through a crawling web bot that organized the web pages. Accidentally, Yahoo! was a bit different. It had humanoid labor of surfing the internet to obtain the desired outcome. Employees in Yahoo created/organized directory of Internet data/content by cut and paste scenario. This procedure managed everything on the web for about two years until the amount of web content exceeded the human limit to maintain a correct structure/format. Meanwhile, many other enterprises were coming up with their crawlers. Companies like Infoseek, Altavista, and Excite brought their programmatic web-crawlers to access and organize the web. Though the outcomes were pretty usable, unfortunately, they did not meet the expectations. Generally, the user had to harvest for the best match on page three or thirty-three. This was another boom time for Google when it successfully made the better use of web scraping and brought it into lights! It made a perfect ratio of human and computer learning, which resulted in making the internet handy to the masses.

Hence, web scraping walks in through the door! Web scraping is a well-known computer bot that crawls the web to retrieve the specific data and converts it in a usable and structured format. The core of the Internet uses web scraping, starring from e-commerce sites most likely, Flipkart, Amazon who compete with the product’s cost ranging to news websites such as CCN, BuzzFeed collecting the top trending news-stories, even the popular web search engines like Bing and Google, make use of one of the forms of web scraping to harvest and organize the data and bring it into a format that is easily accessible by the users. It laid the groundwork for advanced search engines like Google and Bing, yet many enterprises have tried to outlaw the practice. There are problems with web scraping that have presented it to the supreme court and many more with every coming year. There are incidents where crawlers act in illegal ways, and when businesses do not scrape responsibly and unleash their bots on a site causing an accidental DOS denial of service. Nearly ten years before, innovative technology was designed by Stefan Andresen of Kapow Software, known as "visual web scraping.” It enables nonprogrammers to highlight the preferred content from successfully and further brings it into a usable format, mostly excel file, CSV, or database. Also, this idea was immediately recognized by Chris Crabtree, who placed the foundation of Mozenda [3].

III. Web Scraping Tools

Web scraping tools are specially developed software built to ease the process of data extraction. According to Fernandez [4], “A web scraping tool is a technology solution to extract data from web sites in a quick, efficient and automated manner, offering data in a more structured and easier to use the format.” As scraping the data can be a bit exhausting process if done manually, tools come up with different built-in features so that the user can experience the service effectively [5,6,7,8]. These tools are helpful for anyone who is trying to fetch some data online. All the techniques including web crawling and extraction of any form require correct page resources. When it comes to web scraping first way to implement a web scraper is using libraries, including a programming language of our own choice. Many programming languages offer libraries like Java, Python, Ruby as well as Javascript with the framework of Node.js [9]. These libraries use the HTTP protocol to fetch the HTML pages. Some libraries are curl and wget[10]. Users have different needs when it comes to extracting data, tools offer a wide range of facilities that will help the user boost up their network. The second approach is to use web scraping frameworks [11,12]. They help with the process without needing to integrate all the libraries. A python framework, known as scrapy helps to ease the process by integrating all the libraries in a base Spider class to be used in each process. Some other libraries are Jsoup and Web-Harvest.

If the user is not known to coding different desktop-based software are trendy which provides a GUI based environment and ease the process, where the user only needs to enter the URL and click on the data he wants to extract [13]. These tools will not include X-path queries for fetching the data. Later these tools are divided into two groups, partial and complete. Partial tools are more focused on extracting one particular element and are mostly available as a chrome extension [14,15]. Whereas complete tools are the ones that offer a wide range of facilities also including GUI, data fetching and storage, etc. These tools have their varying features, pros and cons [16,17,18,19, 20, 21]. These tools along with their pros and cons are discussed in the following paragraphs.

A. *Dexi.io:*

Dexi.io is formally known as cloudScrape. CloudScrape establishes a good data collection from multiple websites and demands no download similar to a Web host. It renders a "browser-based editor" to establish crawlers and retrieve data in real-time. the user has the option to protect the collected data on cloud platforms like Google Drive and Box.net or export as CSV or JSON. CloudScrape also encourages unknown data access by allowing a set of proxy servers to hide user integrity. it stores the data on its servers up to two weeks before archiving it. It offers 20 scraping hours for free and does cost $29 per month.

Pros:

- Dexi is a very powerful tool, easy to learn, fast: extract data from any website.
- It has a modular interface that is extensively adaptable to build any custom tool/s required by the user.
- It Supports Visual programming: The preview area displays an ongoing scrapped page with its current state inside the robot. It provides a generic timeline interface to build and visualize the robot workflow.
- The provider's support team help to through all thick and thins. The user is just required to Sign up where the browser app opens for the user in the creation of a robot.
- For a commercial tool, the standard plan priced at $119/month (for small projects) is very reasonable and the professional plan would be apt for a larger business need.
- It has an Instinctive interface.
- Supports the Data pipeline, where the output of one element is input for the next one.
- It contains Lots of integration.
- Easy for nonprogrammers to use as No coding is required.
- It supports Agent creation services available.
- It can be easily customized.

Cons:

- The add-ons in Dexi.io that feel attractive at first, do become unmanageable as there is an increase in the cost for any add-on in the store.
- It is quite costly.
- Not very flexible for a layman user.
- It is difficult for non-developers.
- There comes Trouble in Robot Debugging.
- Pretty complex for beginners.

B. *Import.io:*

Import.io grants a builder to produce its own "datasets" by introducing the data from a specific webpage and then further exporting the data to CSV structured format. the user can simply scrape hundreds to thousands of web pages within minutes without any code and the tool can build 1000+ APIs based on the user requirements.

Import.io utilizes "cutting-edge technology" to retrieve millions of data each day, which industries can avail for modest fees. Along with the web tool, it also contributes to free apps for Windows, Mac OS X and Linux to build data extractors and crawlers, by downloading data and syncing with the online account

Pros:

- It has been one of the best UI.
- It is very easy to use and support almost every system.
- It has a user-friendly, clean interface and simple dashboard
- No coding is required.
- A has a generic and light-weight User Interface that works well for non-programmers, who are looking to build their list of track price changes.
- It's a reliable option for scraping at a reasonable speed efficiently from different websites concurrently.

Cons:

- Each sub-page costs credit.
- The tool is self-serve, meaning you won't get much help if you have problems with it.
- Just as lots of another visual web scraping tool, it is also expensive. (basic plan begins at $299/month)
- One has to manually enter the URLs.

C. *ParseHub:*

ParseHub helps in crawling one or many websites with the help of JavaScript, AJAX, sessions, cookies, and redirects. The tool makes use of machine learning technology to identify the complicated documents on the web, hence generates the output file based on the required data format.

Apart from the form of the web app, ParseHub is also available as a free desktop application for Windows, Mac OS X, and Linux that offers a basic free plan that covers 5 crawl projects. This service offers a premium plan for $89 per month with support for 20 projects and 10,000 webpages per crawl.

Pros:

- ParseHub has a rich UI and it retrieves data from several complex areas of a website, unlike other scrapers.
- Developers can meet the ParseHub's Restful API for well data access.
- It Export to JSON / CSV files.
- It supports Scheduler (you can choose to execute your scraping task hourly/daily/weekly).
- Successfully support DropBox as well as S3 integration.
- It also supports multiple systems and Data aggregation from multiple websites.
- Also has a web browser extension.

Cons:

- The speed at which scrape is performed needs to be vastly improved which also slows down the rate at which large volume scrape is done.
- It has a Steep learning curve.
- It is an expensive tool.
- Its Free Program is Limited.
- It has a complex user interface.

D. Mozenda:

Mozenda allows a "cloud-based" web scraping service, alike Octoparse cloud extraction. It is one of the "oldest" web scraping software in the market. It works with a high level of flexibility, has user-friendly UI with all the basic requirements needed during the start of any project. Mozenda comes in two different parts: "the Mozenda Web Console "and "Agent Builder". The Mozenda agent builder is a Windows application practiced for developing a scraping project whereas the web console is a "web application" providing users to set schedules to run the projects or access to the retrieved data. Quite Related to Octoparse, Mozenda also relies on a Windows system and can be a little complex for Mac users.

Pros:

- It is great for big companies and can be integrated into any system.
- It can even scrape PDFs easily.
- It has a visual interface and a Comprehensive Action Bar.
- It provides Multi-threaded extraction and smart data aggregation.
- You can export data into a cloud storage provider such as draw box
- Instability issues in extra-large websites.

Cons:

- It turns out to be unstable when dealing with large websites.
- It is a bit expensive.

E. Octoparse:

It is a Powerful web scraper with comprehensive features for extracting all types of data from the website. It stimulates the human operation process to interact with the website as a result flow of extraction is simple easy and smooth. Using this tool, one can export the data as HTML, CSV. No coding is required to implement Octoparse. It has some predefined templates for amazon and TripAdvisor. Tools like XPath, RegEx, databaseAutoExport, API can be worked on without any extra coding. Octoparse offers its service for both static as well as dynamic websites. It is possible to run an extraction project either on your local machines (Local Extraction) or in the cloud (Cloud Extraction). Octoparse provides a visual operation pane, which is very user-friendly and straightforward. It simulates human web browsing behavior like opening a web page, logging into an account, entering text, pointing-and-clicking the web element, etc.

Pros:

- It has moderate pricing.
- It supports cataloged crawling features and has an outline for unlimited web pages per crawl that has made make it an excellent choice for price monitoring projects.
- Features provided in their free plan are more than enough if someone is looking for an effective one-time tool.
- The precise extraction of data can be achieved with their in-built XPath and Regex tools.

Cons:

- Octoparse is yet to add pdf-data extraction and image extraction features (just image URL is fetched) so calling it a complete web data extraction tool would be a tall claim.
- It is operable in Windows only.

All these tools can be compared using the following tabular representation, ie table 1.

TABLE I. COMPARISON OF VARIOUS WEB SCRAPING TOOLS

Name of the tool	Cost	Complexity	Type	Output	Tools supported
Dexi.io	Free for higher plan $99/ month	User friendly	Browser-based	Excel spreadsheet	CSS3
Import.io	$249, $399 and $799 for different plans.	Difficult for beginners	Desktop application	CSV	XPath
ParseHub	Free, charge for premium plans.	User-friendly	Desktop Application	CSV and JSON	XPath, CSS
Mozenda	First 30 days free (charged for Premium services)	Robust and efficient	Desktop and Web Console	CSV, TSV, or XML	XPath, API
Octoparse	Fourteen days of a free trial.	User-friendly	Both Browser and Desktop	XLS,JSON, CSV,HTML	XPath, RegEx, AutoExport

IV. Conclusion

Among all the renowned paradigms of today's era, one is Web Scraping. There are many tools available for web scraping in the market. By studying the most powerful web scraping tools we have compared the tools in terms of their advantages and drawbacks. We have also discovered certain loopholes that each tool had. To make the tool an ideal fit for web scraping, it is required that the tool offers a wide range of services along with more storage formats and less complexity.

REFERENCES:

[1] V. Shrivastava, "A Methodical Study of Web Crawler" Journal of Engineering Research and Application, vol. 8, issue. 11, pp. 01-08.

[2] A. V. Saurkar, K. G. Pathare, and S. A. Gode, "An overview on web scraping techniques and tools," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, issue. 4, pp. 363-367.

[3] T. Karthikeyan, K. Sekaran, D. Ranjith, and J.M. Balajee, "Personalized content extraction and text classification using effective web scraping techniques." International Journal of Web Portals, vol. 11, issue. 2, pp.41-52.

[4] O. C. Fernandez, "Web Scraping: Applications and Tools." European Public Sector Information Platform, Topic Report.

[5] D. Glez-Pena, A. Lourenco, H. Lopez-Fernandez, M. Reboiro-Jato and F. Fdez-Riverola, "Web scraping technologies in an API world", Briefings in bioinformatics, vol. 15, issue. 5, pp. 788-797.

[6] D. S. Sirisuriya, "A comparative study on web scraping."

[7] A. V. Saurkar, K.G. Pathare, and S. A. Gode, "An Overview On Web Scraping Techniques And Tools," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, issue. 4, pp. 363-367.

[8] Matta, P., N. Sharma, D. Sharma, B. Pant, and S. Sharma. "Web scraping: Applications and scraping tools." International Journal of Advanced Trends in Computer Science and Engineering 9, no. 5 (2020): 8202-8206

[9] P. Raulamo-Jurvanen, K. Kakkonen, and M. Mantyla, "Using surveys and web-scraping to select tools for software testing consultancy," in International Conference on Product-Focused Software Process Improvement pp. 285-300. Springer, Cham.

[10] A. Herrouz, C. Khentout, and M. Djoudi, "Overview of web content mining" tools. arXiv preprint arXiv:1307. 1024.Zhao, Bo. "Web scraping." In Encyclopedia of big data, pp. 1-3. Springer, 2017.

[11] V. Krotov and L. Silva, "Legality and ethics of web scraping."

[12] V. Krotov and M. Tennyson, "Research Note: Scraping Financial Data from the Web Using the R Language. Journal of Emerging Technologies in Accounting," vol. 15, issue. 1, pp. 169-181.

[13] R.S. Chaulagain, S. Pandey, S. R. Basnet, and S. Shakya. "Cloud-based web scraping for big data applications." in 2017 IEEE International Conference on Smart Cloud (SmartCloud), pp. 138-143. IEEE, 2017.

[14] A. Bradley, and R. J. James, "Web scraping using R. Advances in Methods and Practices in Psychological Science," vol. 2, issue. 3, pp. 264-270.

[15] G. Grasso, T. Furche, and C. Schallhart, "Effective web scraping with oxpath." in Proceedings of the 22nd International Conference on World Wide Web (pp. 23-26).

[16] https://www.getapp.com/business-intelligence-analytics-software/a/dexi-io/reviews/

[17] https://www.kdnuggets.com/2018/07/ultimate-list-web-scraping-tools-software.html

[18] Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M. and Mbaye, S.N., 2019, December. Web scraping: state-of-the-art and areas of application. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 6040-6042). IEEE.

[19] K. L. Anglin, "Gather-narrow-extract: A framework for studying local policy variation using web-scraping and natural language processing." Journal of Research on Educational Effectiveness, vol. 12, issue. 4, pp.685-706.

[20] B. G. Dastidar, D. Banerjee, and S. Sengupta, "An Intelligent Survey of Personalized Information Retrieval using Web Scraper." IJ Educ. Manag. Eng.

[21] P. Juyal and S. Sharma, "Locating people in Real-World for Assisting Crowd Behaviour Analysis Using SSD and Deep SORT Algorithm," *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2021, pp. 350-353