ISSUES AND CHALLENGES OF WEB SCRAPING: HEALTHCARE INDUSTRY CASE STUDY APPROACH

Prof. Shravani Pawar Assistant. Professor,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai,
pawarshravani81@gmail.com

Dr. Priya Chandran Assistant. Professor,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai,
priyaci2005@gmail.com

Mr. Pawan Salvi, Student,
Master of Computer Application,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai
, pawanpsalvi2610@gmail.com

ABSTRACT
The Internet is the greatest source of information mankind has ever created. There are various distinct materials in many formats including audio,video, text, etc. However, the data that makes up much of the internet is disorganised, making it challenging to extract and use in automated procedures. Web scraping avoided this labour-intensive manual process of organizing and extracting information, providing an easy way to collect facts and figures from web pages, transform it to a format of your choice, and store it locally. Web scraping has a wide range of uses, which includes brand monitoring, sentiment analysis and data augmentation. Many organizations use different methods to extract useful information. This research paper focuses on various web scraping tools and libraries that have been developed in recent years and are widely utilised to gather data, transform it into structured data, and utilise this organised data in word processing applications. We also discussed the issues and challenges of implementing web scraping techniques in the healthcare industry.
Keywords: Web Scraping, Web data, Data extraction, Data analysis, Extraction error handling

Introduction

Most computer users use the internet and browse the websites using multiple browsers, where data and multimedia are displayed in a way which is easy to understand. In spite of the fact that doing so is totally up to the site owner, many websites provide APIs that can be used to swiftly access much of this data. It is straightforward for them to decide not to grant API access to this data. On the other hand, web scraping and web crawling are faster and more viable methods that can be used to collect information from thousands or even millions of web pages. This technique is pretty beneficial for a variety of applications, however it excels in commercial enterprise intelligence. Since it facilitates them to make decisions, information is critical for groups and organisations, specifically for the reason that the majority of data is now available online. Data collection from many sources, including both public and private ones, is the primary step in any data science research or development. Company sales data and financial reports are examples of private sources. Open data, websites, and journals are examples of public sources. Website analysis, website crawling, and data organisation are the three key, related stages in online scraping. Web scraping and data mining are distinct from one another since the latter includes data analysis while the former is not relevant in this situation. For data mining, sophisticated statistical techniques must also be applied. Because there are so many widely available tools and libraries that offer productive executions of a substantial chunk of the required functionality, web scraping is frequently a very simple process. The capacity to send distinct HTTP requests with varying headers and payloads is a unique feature of the majority of web scraping programmes. This study examines online scraping, including what it is, working, technologies used and how it connects to business intelligence and artificial intelligence. We have also discussed the issues and challenges of implementing web scraping techniques in the healthcare industry.

Literature Review

Ferrara (2014) discussed a number of technological challenges relating to the amount, diversity, velocity, and genuineness of data on the web has to be resolved before web data may be utilised. Quantitative and subjective information are exchanged in a variety of organised, semi-structured, and unstructured formats on the web, including web pages, HTML tables, web databases, emails, tweets, blog posts, photographs, and videos.

Glez (2014) studied web scraping as the automated extraction and structuring of data from the web using technological tools with the intention of further analysing this data. Individual researchers or even big study teams would find it difficult to physically collect and compose Big Data that is readily available on the Web because of its amount, diversity, velocity, and authenticity.

Baumgartner (2005) studied website analysis, website crawling, and data organising are the three main, interwoven stages of online scraping. Website analysis is looking into a website or Web repository's fundamental structure in order to comprehend how the necessary data is kept. This calls for a fundamental comprehension of the World Wide Web architecture, mark-up languages (such as HTML, CSS, XML, and XBRL), and various Web databases (e.g. MySQL). A script which automatically browses a website and obtains the necessary information is created and run as part of website crawling.

Fernandez (2011) discussed the availability of tools for the automated crawling and parsing of web data has something to do with the common request of these languages in Data Science. The techno-logical outline of business analytics accordingly supports analytics implanted in decision support activities of businesses. As a holistic approach, business analytics encompasses all disciplines of business administration.

Hillen (2019) It is vital to clean, pre-process, and arrange the required data after it has been extracted from the chosen web source. This will allow for further analysis of the data. Given the amount of data involved, a programmed approach could also be required to reduce the amount of time spent. Natural Language Processing (NLP) libraries and data manipulation methods are available in several computer languages, including R and Python, and are helpful for cleaning and organising data.

Zhou (2014) in his study the authors have focused on extracting web contents which are less structured, such as new articles. They also have proposed a method to automatically cluster and extract based on the relevance identified.

Michalakidis (2016) proposed different techniques to extract data from different structured and unstructured sources. The authors have proposed an error reporting mechanism associated with collection of data and metaphysical based data dictionaries for improving the quality.

Kumar (2020) stated that data mining techniques are extensively used in healthcare research. Most of the data collected are unstructured and it is very difficult to collect that data manually. The authors have discussed web scraping techniques to collect data from unstructured HTML documents and store it in a format capable of doing data analysis.

Singrodia, Khder & Krotov (2022) various aspects of web scraping and its tools and techniques are discussed the legal and ethical issues related to web scraping techniques used.

Hillen (2019) discusses how web scraping can be used in food price research. He also discusses data collection methods on different real time data. The authors have also discussed the limitations in terms of non-availability dataset in food price research.

Tools & Techniques used in Web Scraping

Python:

Python is an object-oriented, high-level programming language. Code clarity is prioritised in its layout philosophy, which makes heavy use of indentation. It supports quite a few programming paradigms, which include procedural, object-oriented, and practical programming in addition to established programming.

Beautiful Soup:

A Python module alluded to as Beautiful Soup is utilised to parse HTML and XML texts. For processed pages, it produces a parse tree that can be utilised to extract HTML facts for net scraping. The application is likewise funded with the aid of using Tide lift, a paid open-supply preservation subscription.

Requests:

One essential aspect of Python for sending HTTP requests to a given URL is the requests module. REST APIs and internet scraping each want requests, which needs to be learnt earlier than the use of those technologies further. A URL responds to requests by returning a response. Python requests have integrated control equipment for each request and the response. Python customers may also put up HTTP/1.1 requests and the use of the HTTP library requests, which is licenced beneath the Apache 2 licence. Python requests are essential to test with the internet. One needs to ship a request to the URL so that you can perform moves including hitting APIs, downloading complete Facebook pages, and performing many different things

JavaScript:
The lightweight object-oriented programming language JavaScript (is) is used by many websites to script their webpages. It is a full-fledged, interpreted programming language that, when combined with an HTML content, enables dynamic interactivity in websites.

Discussion and Analysis
Broadly speaking, web data scraping is a method of systematically collecting and combining information from different web sources. A software agent known as a web robot simulates the browsing interaction between web servers and the users. The robot systematically scans a required number of websites, analyses their contents to identify and extract relevant information, and then organizes that content as needed. Web scraping APIs and frameworks are commonly used by the organizations to extract the useful information.

We identify the tools that are now available on the market, explain the terms "web scraping" and "web crawling," and instruct readers on how to build their own web scrapers using one of these tools. The Web data scraper connects to the target Web site via the HTTP protocol, a stateless text-based Internet protocol that governs request-response interactions between a client, frequently a Web browser, and a Web server. Web data scrapers must carefully arrange their retrieval tasks to avoid overloading the server and must abide by the site's terms of usage. After obtaining the HTML file, the Web data scraper can extract the desired contents. For this reason, regular expression matching is widely employed, either independently or in conjunction with other justifications.

Web scraping process is depicted in Figure1. It involves amassing all of the records that have been retrieved from a couple of URLs provided, which can be structured, semi-structured or unstructured. The information is extracted and it should be pre-processed, cleaned and transformed to a format which can be used according to the business requirements. After the transformation, the records are stored inside the business database for further usage.

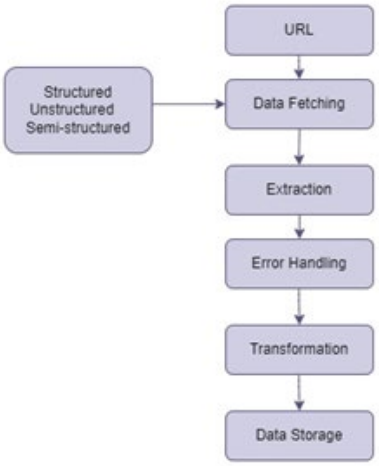

Fig1: Web Scraping Process

Features of web Scraping
1) Pace
The speed that web scraping technology offers is by far its most prominent advantage. Web scraping enables you to quickly rub a few pages at once without having to screen and oversee each individual request. Web scraping's speed is due in part to its ability to scan web pages quickly and extract data from them as well as the ease with which it may be incorporated into daily activities. Because you don't need to worry about creating, downloading, integrating, or installing web scrapers, getting started with them is simple.

2) Profitability
Web scraping offers a complex benefit at a sensible cost, which is one of its best highlights. You won't have to contribute to developing a complex system because a basic scraper can frequently complete the entire task. A professional data extraction project would be impossible without automation because time is money and the web is developing at an accelerated rate.

3) Flexibility and methodical approach
Web scraping programmes and APIs are not pre-programmed fixes. They are therefore very open, adaptable, and interoperable with other scripts. Make a scraper for a single, large work, then modify it to meet a variety of smaller tasks by making just minor changes to the core. For instance, a workflow made of various APIs may be used to scrape every monitor offered by Tata 1mg and compare it to the ones you sell online. The online scraping API gives users the ability to personalise the data gathering and analysis process and make the most of its capabilities to achieve all of their web scraping goals.
4) Performance reliability

Data accuracy is a process in and of itself that web scraping offers. Once properly configured, your scraper will accurately and reliably capture data directly from websites with very little risk of error. It is crucial to have data in a comprehensible and orderly manner in addition to being able to gather it. If the script is properly written, you can virtually remove the possibility of error and guarantee that the information and data you obtain are of higher quality each and every time you gather them.

5) Automatic delivery of structured data
Simple values may frequently be utilised right away in other databases and applications because of the fact that well-scraped data always comes in a machine-readable format by default. Its most appealing feature for both professionals and non-pros is the simple API interaction with other applications. The initial stage in your data analysis pipeline that includes other built-in solutions is web scraping. Web scraping is a difficult operation since it involves many different computer languages and software programmes. The greatest thing is that a web scrape will also do the required account maintenance, such as simple troubleshooting, updates, and backups. I can be confident that my data is secure when I use web scraping, which is the main advantage.

Limitations
1) Needs perpetual maintenance
Maintenance of the product is the actual deal. You have no influence over whether an external website changes its HTML layout or content because your scraper's activity is inextricably linked to it. Developers must therefore respond to those changes in order to prevent scrapers from breaking or becoming out of date and unable to keep up. While some information will be updated automatically, any scraper will often require ongoing maintenance to remain operational.
2) Data Extraction
Setting realistic expectations is crucial when working with complex data extraction and processing. The main purpose of a scraper is to gather the necessary sort of data, package it in the format you require, and upload it without loss into your computer or database. Although the data will be delivered in a structured format, more complicated data will need to be processed before it can be incorporated into other applications.
3) Scrapers can get blocked
Some websites simply dislike being scraped. They might do this because they think scrapers are using up their resources, or it may essentially be that they do not need to create it basic to match businesses to compete. Access can occasionally be denied due to the scraper's place of origin. The usage of proxy servers is a common solution to this form of IP blocking. In many situations, these techniques can let scraping bots work covertly. In any case there are circumstances when even these remedies are inadequate to handle severe blocking, and the final drawback is that a website cannot be scrapped.
4) Learning Curve
It takes practice to master even the simplest scraping tool. Some tools still need you to know how to code. Some tools for non-coders may take weeks to learn. Understanding of XPath, HTML, and AJAX is required in order to efficiently scrape webpages.
5) The structure of websites change frequently
Data that has been scraped is organised in accordance with the website's structure. Sometimes when you visit a site again, the design has changed. Some website designers update their work frequently to improve user interface, while others may do so to prevent scraping. Changes to the website layout might be significant or little, such as moving a button's position. Your data can become corrupt even with a small alteration. You must modify your crawlers every few weeks in order to obtain accurate data because the scrapers were constructed using the old site as a guide.
6) Data extraction on huge scale is difficult
Some tools can only handle small-scale scraping, therefore they cannot extract millions of data. Owners of ecommerce businesses that require millions of lines of frequent data feeds directly into their database are bothered by this. Multiple cloud servers are used to perform tasks. You get blazing speed and colossal storage space.
7) A web scraping tool is not omnipotent

Texts may be extracted from source code and formatted using regular expressions using sophisticated programmes. Pictures can only be scraped for their URLs, which may then be transformed into images. It is significant to highlight that because most web scrapers gather data by parsing via HTML components, they are unable to crawl PDFs.

Case Study Approach: Web scraping in Health Sector

In our research we have conducted a study on the challenges and issues of web scraping in the healthcare industry. The fact that public health records vary in size and type is undeniable. Web scraping too is slowly and steadily gaining importance in healthcare. The truth that, vast amount of information produced by the medical industry is very difficult to analyse using traditional methods. So, web scraping along with data mining can improve decision making by identifying patterns and trends in large volumes of complex data. One can filter some websites with web scraping and use the information for public health analysis, prescription drugs pricing analysis, disease monitoring, insurance database, competitive analysis etc. In this case study first we discuss the data extraction problems in the healthcare industry. Most of the healthcare research depends on the data collected from Electronic Patient record (EPR) and data repositories. Most of the EPR data is unstructured or semi structured (Michalakidis G, 2016). Also these data are collected from heterogeneous sources and each may have its own structures. These problems of collecting data from heterogeneous sources can be concluded as,

- Local autonomy
- Architectural difference
- Representational dissimilarity
- No Precise interpretation

Generic approach for error reporting.

The above problems can be avoided if the data is collected from a single source. Since the data is very crucial for the research we cannot restrict to data from a single source. The solution for this is to the need for a structured or generic approach for data extraction. By using this approach, we can categorize the extraction error into different classifications and can address these issues according to the severity of the error. This paves the way for creating an online generic approach for error reporting.

Conclusion

Online scraping is a well-known term that has gained more prominence as a result of the need for free data gathered from web pages. The data are needed by many professionals and researchers for processing, analysis, and the extraction of important outcomes. In contrast, those working with use cases need to allow data from many sources to be integrated into creative applications that will provide supplemental benefits and originality. We have examined the many facets of web scraping, starting with the web scraping tools and software, and also looked at their advantages and disadvantages. Then discussed the challenges and issues of web scraping in the healthcare industry. We look forward to extend our study in other sectors like the investment sector and also to study novel approaches for data collection using web scraping.

References

Baumgartner, R., Fröhlich, O., Gottlob, G., Harz, P., Herzog, M., & Lehmann, P. (2005). Web data extraction for business intelligence: the lixto approach. Gesellschaft für Informatik eV.

Fernández V, J. I., Blasco Garcia, J., Iglesias Fernandez, C. A., & Garijo Ayestaran, M. (2011). A semantic scraping model for web resources-Applying linked data to web page screen scraping.

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-based systems, 70, 301-323.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.

Hillen, J. (2019). Web scraping for food price research. British Food Journal, 121(12), 3350-3361.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing & Its Applications, 13(3).

Krotov, V., & Johnson, L. (2022). Big web data: Challenges related to data, technology, legality, and ethics. Business Horizons.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping.

Kumar, V., Thareja, R., Thareja, R., & Jain, P. R. (2020). Applying Data Science Solutions in the Healthcare Industry. In ICT for Competitive Strategies (pp. 35-42). CRC Press.

Michalakidis, G. (2016). Appreciation of structured and unstructured content to aid decision making-from Web scraping to ontologies and data dictionaries in healthcare. University of Surrey.

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scraping and its applications. In 2019 international conference on computer communication and informatics (ICCCI) (pp. 1-6). IEEE.
Zhou, Z., & Mashuq, M. (2014). Web content extraction through machine learning. Stanford Univ, 1-5.

