DOI: 10.15849/IJASCA.211128.11

Int. J. Advance Soft Compu. Appl, Vol. 13, No. 3, November 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application

Teilweise gelesen

Moaiad Ahmad Khder

Department of Computer Science, College of Arts and Science, Applied Science University, Bahrain
e-mail: moaiad.khder@asu.edu.bh

Abstract

Web scraping or web crawling refers to the procedure of automatic extraction of data from websites using software. It is a process that is particularly important in fields such as Business Intelligence in the modern age. Web scrapping is a technology that allow us to extract structured data from text such as HTML. Web scrapping is extremely useful in situations where data isn't provided in machine readable format such as JSON or XML. The use of web scrapping to gather data allows us to gather prices in near real time from retail store sites and provide further details, web scrapping can also be used to gather intelligence of illicit businesses such as drug marketplaces in the darknet to provide law enforcement and researchers valuable data such as drug prices and varieties that would be unavailable with conventional methods. It has been found that using a web scraping program would yield data that is far more thorough, accurate, and consistent than manual entry. Based on the result it has been concluded that Web scraping is a highly useful tool in the information age, and an essential one in the modern fields. Multiple technologies are required to implement web scrapping properly such as spidering and pattern matching which are discussed. This paper is looking into what web scraping is, how it works, web scraping stages, technologies, how it relates to Business Intelligence, artificial intelligence, data science, big data, cyber security، how it can be done with the Python language, some of the main benefits of web scraping, and what the future of web scraping may look like, and a special degree of emphasis is placed on highlighting the ethical and legal issues.

Keywords: *Web Scraping, Web Crawling, Python Language, Business Intelligence, Data Science, Artificial Intelligence, Big Data, Cloud Computing, Cybersecurity, legal, ethical.*

1 Introduction

For the vast majority of computer users, interaction with the internet comes in the form of accessing websites through a browser, where information and multimedia objects are displayed in a human-friendly fashion for easy understanding. However, this is only the tip of the iceberg in regards to the potential of the web, as there is so much useful raw information that is hidden out of view. APIs can be used across many sites to easily access much of this information, but these are at the discretion of the site owner, and should they choose not to make this information accessible through APIs, it’s easy not to. Web scraping or web crawling on the other hand is far faster and more effective, and can be used to gather and compile data from across thousands, or over millions, of pages for processing and drawing information from. This technique is invaluable across a wide variety of applications, but in particular in the field of Business Intelligence. Any company wishing to keep up in the 21st century needs to have a strong online presence, and one important tool for maintaining that while keeping up with competitors is web scraping. (Lawson, 2015).

Data is very important for businesses and organizations as it assists their decision making and especially, currently, most of the data can be found on the internet. (Almaqbali *et al*, 2019). Data acquisition is the first phase in any data science study and development; it is the step in which the data obtained from private sources such as firm sales records and financial reports, or from public sources such as journals, websites, and open data, or by purchasing data. (Chaulagain *et al*, 2017). Website analysis, website crawling, and data organizing are the three primary, interwoven processes of online scraping. (Milev, 2017) Web scraping differs from data mining in that the latter entails data analysis, whereas collecting data is immaterial in this situation. Data mining also necessitates the use of sophisticated statistical techniques. (Krotov and Silva, 2018). Due to the wide number of accessible tools and libraries that offer efficient implementations of much of the required functionality, web scraping is a pretty simple process in general. The ability to send custom HTTP requests with different headers and payloads is standard feature of most web scraping programs. (Yannikos *et al*, 2019).

This paper is looking into what web scraping is, how it works, web scraping stages, technologies, how it relates to Business Intelligence, artificial intelligence, data science, big data, cyber securityو how it can be done with the Python language, some of the main benefits of web scraping, and what the future of web scraping may look like, and a special degree of emphasis is placed on highlighting the ethical and legal issues

2 Related Work

The popularity of the Internet led to rapid increase in amount of data created each day, which prompted the need to scrape websites. Search engine engineers created the first well recognized scrapers like Google. These scrapers search over the whole Internet, scanning each web page, extracting information, and compiling a searchable index. (Lawson, 2015)

Web scraping/web crawling is extracting information from website through the computer software, this software can will copy the way the human explore the world wide web, by executing a fully-fledged web browser, like Mozilla Firefox or internet explorer, also another way to copy the way the human the world wide web, it through the low-level Hypertext Transfer Protocol (HTTP). In other words, instead of copy-paste information manually from the website and gathering it in a spreadsheet, web scraping offers this functionality in computer applications that can accomplish it more precisely and much quicker than a human. (Lawson, 2015)

Web scraping is a technique for converting unstructured web data into structured data that can be saved and analyzed in a central spreadsheet or database. This enables the bot to retrieve large volumes of data in short amount of time, which is advantageous in today world especially since we have big data which is always changing and updating. For example, assume you have a clothing shop and you want to keep track of competitor cloth pricing. You can go to competitor website and gather information every day in order to compare them with your pricing for each product, but this will take a lot of time. Furthermore, if you have thousands of competitors it will be very difficult to keep track of the pricing this is where web scraping comes in handy. (Banerjee, 2014)

Web scraping is a technique for converting unstructured web data into structured data that can be stored and analyzed in a central database or spreadsheet (Sirisuriya, 2015). Web scraping is used in different industries such as cyber security, cyber security, etc. it can be used to determine prices and costs of production by assessing the data that has been gathered. Ads can be created to advertise products to different users depending on their data such as location and cookie settings. It can be relative described as a new method for data collection on the internet. Web scraping is already being used for scientific and commercial applications since it allows for the development of big, customized data sets at minimal costs. It is meant to replace outdated methods of data collection as more business are looking to catch up on new trends that are followed by consumers. (Hillen, 2019).

Web scraping API which is a web scraping service that is in larger scale, based on customized requests web scraping. It provides institutions with structured data by accessing to scraped data of their clients using API. (Mitchell, 2018)

Web scraping was the only way for a program to obtain information from the internet until the recent emergence of APIs, which are special interface designed to make communication easier for applications and servers. API are incredible tools which can provide the data in an organized way. API can be access for variety of different data types like Wikipedia article, tweets in twitter and many more. Even though some website offers API, but not all API are free to use and also, they are very limited in terms of how much data they provide and what data they provide. Furthermore, website developer's priority is to focus and maintain the frontend interface over backend API. To summarize, we cannot depend on APIs for acquire the internet data we need, so we need to apply web scraping strategies in order to ensure that the data we are getting matches with what the business or company needs. (Mitchell, 2018) (Broucke and Baesens, 2018).

2.1 Web Scraping Concepts and Techniques

Data is extracted from sites using the HTTP protocol used by web browsers. This process can be done with manual browsing or automated with web crawlers. Webs scrapping is one the most valuable tools available to data scientist as it allows the extraction of huge amount of data that is constantly generated online with a relatively low cost. (Zhao, 2017). Proper data preprocessing and cleanup is required for proper utilization of scraped datasets (Tarannum, 2019) discuss cleanup methods for scrapped data in python. (Manjushree and Sharvani, 2020) Perform a review of web scrapping technology from literature and discuss tools, methodology and process for web scrapping they also discuss the fields and targets of web scrapping. (Grasso *et al*, 2013) Introduce a new tool and language for web scrapping called OXPath, OXPath is an extension of the XPATH standard which adds several features such as user actions such as clicks and filtering by visual features exposed from CSS such as color. They also discuss several projects utilizing OXPATH such as DIADEM an unsupervised data extraction tool. (Gheorghe *et al*, 2018) discuss modern techniques and challenges in web scrapping such as captcha, rate limiting, and they discuss a case study of scrapping public transportation data from Bucharest Public Transportation authority website to analyze issues in public transportation.

Traditional copy and paste technique: The most effective and practical web scraping technique is usually a copy-and-paste and manual analysis technique. However, when users need to scrap a large number of datasets, this is a mistake, tedious, and unpleasant procedure. (Sirisuriya, 2015)

Grabbing text and using regular expressions: This is a significant and simple method for extracting data from web pages. This strategy is based on the UNIX command or the computer language's regular expression-matching capabilities. (Sirisuriya, 2015)

Programming with the HTTP: Users can access data in both static and dynamic webpages applying this method. Data may be retrieved using socket programming via making HTTP requests to a remote web server. (Sirisuriya, 2015)

HTML parsing: semi-structured data query languages is used for parsing webpages HTML code and retrieving and transforming page content

DOM parsing: embedding a full-featured web browser, like: Mozilla browser or Internet Explorer control, programs may access dynamic material created by client-side scripts. These browser controls also parse webpages into a Document Object Model tree, from which applications may get some pages contents. (Saurkar *et al*, 2018)

Web Scraping Software: Nowadays several tools that can provide a custom web scraping. These programs can identify page data structure automatically or give a recording interface which eliminates the need for web scraping scripts. Furthermore, some of these software's can have a scripting function which can be used for extracting and transforming material, as well as database interfaces for scraping data and storing it in local databases. (Saurkar *et al*, 2018)

Computer vision-based web page analyzers: By visually scanning web pages like a person, computer vision and machine learning are being utilized to discover and retrieve important information. (Saurkar *et al*, 2018). A very simple illustration for web scraping is shown in fig 1.

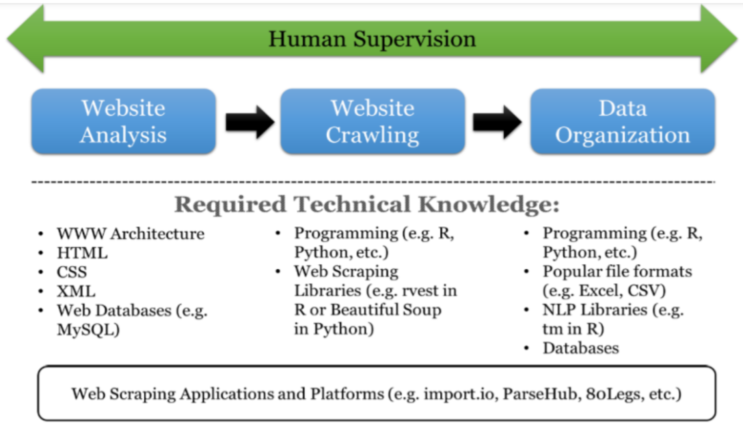

Human Supervision
Website Analysis
Website Crawling
Data Organization
Required Technical Knowledge:
• WWW Architecture
• HTML
• CSS
• XML
• Web Databases (e.g. MySQL)
• Programming (e.g. R, Python, etc.)
• Web Scraping Libraries (e.g. rvest in R or Beautiful Soup in Python)
• Programming (e.g. R, Python, etc.)
• Popular file formats (e.g. Excel, CSV)
• NLP Libraries (e.g. tm in R)
• Databases
Web Scraping Applications and Platforms (e.g. import.io, ParseHub, 80Legs, etc.)

Fig 1: Web Scraping (Krotov and Tennyson 2018)

2.2 Web Scraping Process

The web scraping process is divided into 3 stages as shown in fig.2, which are:

Fetching stage: The desired website with the relevant information must first be accessed in what is known as the fetching phase this is accomplished via the HTTP protocol, which is an Internet protocol for sending and receiving requests from web servers. Web browsers utilize similar methods to get material on web pages. In this step, libraries such as curl 2 and wget 3 can be used by sending an HTTP GET request to the target address (URL) and get the HTML page as a response (Persson, 2019)

Extraction stage: After retrieving the HTML page, the important data should be extracted. Regular expressions, HTML parsing libraries, and XPath queries are utilized in this step, which is referred to as the extraction stage. The XML Path Language (XPath) is a tool for finding information in documents. This is the second phase of the project. (Persson, 2019)

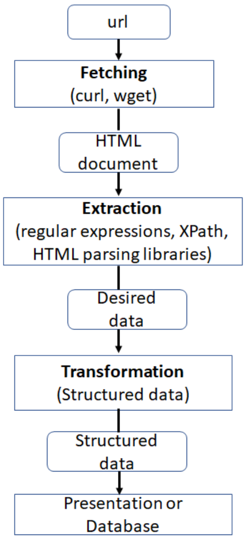


Fig.2 Web Scraping process (Persson, 2019).

Transformation stage: Now that just the relevant data remains, it may be converted into a structured format for presentation or storage. Using the stored data, we can gather information which can help the business intelligence in making a better decision and much more. (Persson, 2019)

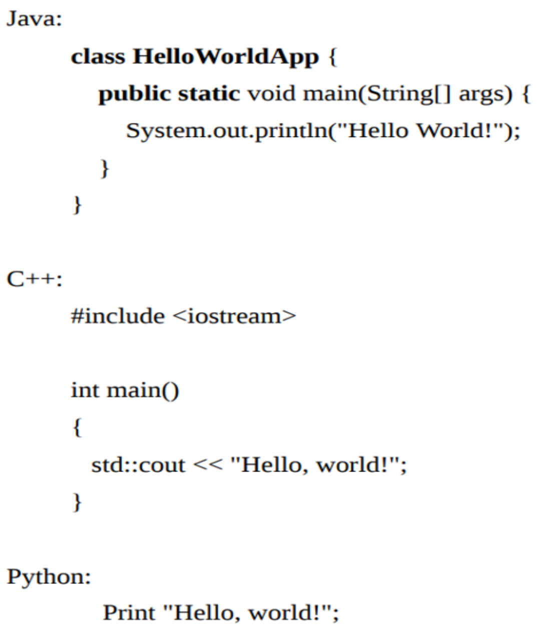
2.2.1 Python for Web Scraping

One of the reasons why python is a good choice for web scraping is that python is one of the most popular language, some people might not hear about python programming language but it’s very easy to learn even if you have not used it before or have not programing experience.

Since python is a popular language it means that if you have a huge community that can help you in case you face any problem or issue, if you have an issue during your programming, it's pretty likely that someone else has had the same problem and solved it somewhere online. This isn't exclusive to Python, but its accessibility and long-standing popularity have resulted in one of the largest and most diversified user communities among today's main programming languages. (CreateSpace Independent Publishing Platform, 2015)

Another reason python is a fantastic choice for web scraping (or learning any programming skill, for that matter) is that Python code is very simple to read. Consider the following example in 3 different programing language. (CreateSpace Independent Publishing Platform, 2015)

Writing a readable code is important not only because it’s easier to keep track while you are working on it but also it allows other programmers to understand what you accomplished and it's much easier to figure out what you were thinking when you wrote it after a few weeks or years. This is why well-written Python code is extremely easy to maintain and reuse. (CreateSpace Independent Publishing Platform, 2015). Very simple illustration for the same code written in 3 different languages (Java, C++, and Python) is shown in fig 3, and its noticeable that python is very easy to read,



```
Java:
        class HelloWorldApp {
           public static void main(String[] args) {
              System.out.println("Hello World!");
           }
        }

C++:
        #include <iostream>

        int main()
        {
          std::cout << "Hello, world!";
        }

Python:
           Print "Hello, world!";
```

Fig.3 Example code in different languages (CreateSpace Independent Publishing Platform, 2015).

2.3 Three Approach of Web Scrapping

Regular expression is a python package which was developed for Perl programming language at the beginning. The way python handles regular expression is by using “re” module. Regular expression operates by establishing a pattern which we want to find in a string and after that it scans for any matches of that string. The pattern might appear strange because it contains special characters that change how the pattern is interpreted and material that we want to match (Uzun, 2018)

Beautiful Soup is a Python package that allows you to retrieve structured data from a webpage which was developed by Leonard Richardson and some other developers. Beautiful Soup used for parsing XML and HTML. Furthermore, it is much easier to use when compare with regular expression since it has a fewer step for navigating, examining and updating a parse tree. Beautiful Soup have the ability to automatically convert outcoming document into UTF-8 and incoming document into Unicode, so you don't need to keep track of encodings unless the document does not specify one. (Uzun, 2018)

Lxml is one of Python's fastest and most feature-rich libraries for XML and HTML processing. Furthermore, lxml is a straightforward which make it easy to learn and XPath is supported by lxml for extracting tree content. You may extract content chunks into a list using XPath. Moreover, if you've worked with CSS or XPaths before, you'll have no trouble picking it up. Its raw power and speed have also contributed to its widespread adoption in the industry. Lxml is a very good tool to use for web scraping but some of the people avoid using it because it’s difficult to install in certain computers (Uzun, 2018)

The benefits and drawbacks of each scraping method is highlighted in table.1:

Table.1 comparison of three approaches of web scrapping (Lawson, 2015)

Scraping approach	Ease to install	Performance	Ease of use
Regular expressions	Easy (built-in module)	Fast	Hard
Beautiful Soup	Easy (pure Python)	Slow	Easy
Lxml	Moderately difficult	Fast	Easy

If the limitation of your scraping is download instead of retrieving data from website, then it will not be an issue for you to use a slower approach like beautiful soup. Another good option would be regular expression if you only care about scraping a limited quantity of data and don’t need any further third parties or dependencies. Overall, the best approach is the lxml because it can powerful and quick whereas the other two approach are only effective in specific situation (Lawson, 2015)

2.4 Usage of Web Scrapping

Web scraping techniques allow users to scrape data from various websites into a single database or spreadsheet. As a result, data may be quickly seen and analyzed for future use. (Saurkar *et al*, 2018)

Web scraping involves the creation and implementation of two software programs: a crawler and a scraper. The crawler downloads data from the Internet in a systematic manner; the scraper then extracts important information in its raw form from the downloaded data, codes it, and stores it in a database or file according to a user-defined structure and format. This new file is then evaluated in ways that the initial data presentation on the internet does not allow for. (Farley and Pierotte, 2017)

Previously, the data was only available on web browsers and it could not be copied. Web scraping has made it possible, and it can be done with a few lines of script in a short amount of time (Parikh *et al*, 2018)

Web scrapping is widely used in business and scientific fields.(Hillen, 2019) mention the use of web scrapping for food price research due to the low cost of information retrieval, high frequency and details in the generated set compared to other sets such as ones provided by national and commercial agencies, their method involves identifying the target site, selecting the data to be captured and output format and finally automating the process. (Yannikos *et al*, 2019) discuss the use of web scrapping to extract and analyze darknet marketplaces product supplies and prices, their method involves retrieving the html data through the TOR network, parsing the relevant data into structured form, storage in relational database and finally statistics and charting from stored data. It is used to determine prices of products and expenditure used to spend on advertisements and commercials by businesses. It can recognize trends and user preferences.

3 Application of Web Scraping

Web scraping is used widely in many different computer science fields, especially, the hot trends and news areas such as: Business Intelligence, AI, Data Science, Big Data, Cloud Computing and Cyber Security

3.1 Application of Web Scraping in Business Intelligence

Price tracking, market research, and sentiment analysis are three of the most common uses of web scraping. In e-commerce, price is extremely important. The rivals must be watched from a business standpoint. When prices vary often, manually tracking all rivals' pricing is not a realistic solution. Web scraping serves a function in this situation: it automates the extraction of price information from rivals and can offer up-to-date information from all of them in one document or database that is easily accessible. (Banerjee, 2014)

For a better decision-making process, market research demands extremely precise data. Market analysts and business intelligence professionals across the world rely on high-quality, meaningful data to do their tasks. This makes online scraping a feasible approach for commercial operations including market pricing, market trend analysis, point of entry optimization, research and development, and competition monitoring. (Vording, 2021)

Web scraping can be used in stock market to visualize the price changes over period of time, and social media comments and feeds have been scraped to know the opinion of the public on opinion leaders. (Sirisuriya, 2015)

Web scraping data from social media for sentiment analysis is a common practice. Predicting elections is one use. By analyzing tweets and postings where the candidate's name isn't even necessary, a computer might guess the winning candidate's name. Sentiment recognition algorithms discover patterns that go beyond the post and detect clues. When utilizing web scraping to collect data, more accurate analysis may be conducted than when using aggregated postings (based on hashtags on Twitter, for example). (Vording, 2021).

Web scraping has a multitude of benefits and advantages. Some of these include: more efficient, consistent, and easy than manually performing the same tasks, very low maintenance, very low running costs, online reputation Offer your customers more targeted advertising, provides insights into pricing data, market dynamics, prevailing trends, your competitors’ practices, and the challenges they face, collect public opinion, scratch lead, develop vertical search engines with a specialized focus, obtaining the most precise and timely results that cannot be obtained by humans and fast and precise findings help firms save time, money, and effort, giving them a clear speed-to-market edge over their competition (Kasereka, 2020; Hillen, 2019; Scraping Expert, 2017)

In addition, the web scraping Have a competitive advantage, where the competitors scrape pricing from e-commerce and marketplace websites using automated bot programs. Competitors use this unlawful behavior to get real-time dynamic pricing data in order to undercut product prices and attract price-sensitive customers. Bot programs may be developed to scrape pricing information from a range of rivals' websites, compare them, and update the portal with the best rates in order to attract more consumers. When potential customers search for items on Google, these rivals come up first, at lower pricing than their competitors. (Krotov and Tennyson,2018; Vording, 2021)

Web scraping is technically possible for every online business that creates original material. The site's susceptibility to scraping is determined by one or more of the following four key factors:

1. Competition from similar businesses is on the horizon.
2. The website's popularity in terms of traffic or user interaction.
3. The stuff that is created/changed is unique or essential in some way.
4. Existing security flaws in websites

However, as a business owner, you can't be sure if you're losing data to the competition or keeping it for educational or research purposes. (Banerjee, 2014; Broucke and Baesens, 2018)

3.2 Applications of Web scraping in AI

Data derived from scientific or academic research is vital in the marketing industry. Scholars, market researchers, and academics gather information from a variety of sources in order to better their products and services. Most website authorities forbid users from downloading data from a website and storing it locally. As a result, the user may desire to manually copy data from the website to a local file storage space on their computer. However, completing such a project would be exceedingly hard and time consuming. Due to these limitations, web scraping methods have been developed. Web scraping techniques allow users to collect data from multiple websites and store it in a single database or spreadsheet. As a result, data may be viewed and evaluated for potential future use with little to no effort. Web scraping is a branch of web mining that focuses on gathering data from the web. An important aim is to provide a comprehensive review of the various web scraping techniques and apps available for the extraction of critical website data. In the Information Retrieval, Data Extraction, and Data Mining triangle, Web Mining is still a vital component. Prior to being processed by data mining tools, retrieval and extraction of important information from unstructured data are critical steps in the process. Exploring these strategies further is critical for dealing with the massive amount of data available in the Information Overload Era. Additionally, as the Web continues to emphasize the importance of semantics and information integration, these fields of study become critical for addressing the Web's new future trends (Saurkar *et al*, 2018). The web mining including (data mining, information retrieval and information extraction) is shown in fig 4.

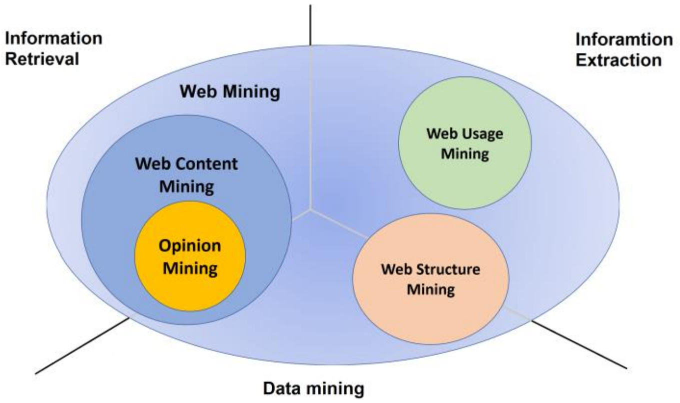


Fig 4: Web Mining (Saurkar *et al*, 2018)

Information Retrieval (IR) is a topic of study that focuses on the document's retrieval from a collection of various documents (both relevant and irrelevant), typically using key word searches. With the evolution of the Internet, Information Retrieval got a new importance level, as search engines became the leading method of locating data on different websites. Nowadays, due to their prominence, search engines are the top used evocative term in the field of information retrieval. The term "Information Extraction" refers to a sub-discipline of artificial intelligence. Internet Explorer (IE) is mostly concerned with extracting valuable info from unstructured data. Typically, an extraction information system is focused on the identification of individuals or objects (people, places, and businesses, for example) and the extraction of rules. Unstructured data includes video, audio, pictures, and text. Initial information extraction systems concentrated mostly on text, which is remains the most investigated sort of information by commercial frameworks and scientific community to this day. The objective of IE is to extract usable components from unstructured data in order to create more value information via semantic classification. The outcome may be applicable to other information processing tasks, such as information retrieval and data mining. While the objectives of IR and IE are fundamentally different, they should be viewed as closely related activities in order to improve their precision and accuracy.

Data has gained extra value as a result of advancements in data computing and analysis technologies. It was difficult to think several years ago that we would ever be able to pull this much information from the Internet. All of this is possible because modern techniques are capable of processing massive amounts of information in a short time. Internet offers access to a vast volume of unstructured or unlabeled material that is difficult for humans to acquire due to their lack of expertise about available information sources. Additionally, many people are unaware that their personal data is available online. The article discusses a system for retrieving personal information from the Internet based on a variety of input criteria. By utilizing artificial intelligence algorithms, the system is capable of differentiating the information of multiple people with the same name. Although the material for performed case study was acquired from different sources holding data about people in Spain, this may easily be applied to other countries' specialized sources of data. This approach was confirmed in a case study including multiple participants, and the findings were extremely satisfactory. (Chamoso *et al*, 2020)

In computing, a bot is an autonomous software that operates on a network (particularly the Internet) and can interact with other systems or users. A description of how an Artificial-Intelligence bot's memory can be optimized through the use of a faster searching algorithm and how it can learn new things that the user desires the bot to learn. A Web-Crawler is a type of bot that crawls through a collection of websites or the entire internet. Web-crawlers are also referred to as Web-spiders. While crawling the entire internet is a too much for a personal assistant, a bot can crawl a few sites and gather the information. Therefore, in order for the bot to gather information from the internet, it must crawl through the websites and scrape the necessary information. To accomplish these duties, a web-crawler or spider is utilized for crawling and a web-scraper is used for scraping. (Bhatia, 2016)

3.3 Application of Web Scraping in Data Science

In domains like Natural Language Processing, Sentiment Analysis, and Machine Learning, retrieving data from social networks is the initial step. Important data science activities rely on historical data to anticipate future outcomes. Most recent works employ Twitter API, a public platform for collecting public streams of information. A new way offered for gathering historical tweets using web scraping techniques that circumvent Twitter API constraints. (Hernandez-Suarez *et al*, 2018).

A good example of using web scraping in data science is retrieving data from social media for different purposes, such as using web scraping of COVID-19 news stories to create datasets for sentiment and emotion analysis (Thota and Elmasri, 2021).

Using real data and teaching data science cycle completely, where this cycle starts with importing data, are best practices in data science and statistics courses. Web is an important source of current information, frequently provided and stored in a format that requires some wrangling and transformation before analysis. In the absence of formal teaching on how to organise data for research projects, students typically resort to manual entry or copy-pasting into a spreadsheet. This process is both mistake prone and time consuming. Web scraping teaching allows for easy integration of such data into the curriculum. Web scraping is explained and how it may be used effectively in statistics and data science curriculum at various levels.

Class activities link this new computer method to classic statistical themes, and finally, a discussion took place for the benefits and drawbacks of web scraping in the classroom, as well as how to avoid them. (Dogucu *et al*., 2021).

Why Should Data Scientists Use Web Scraping? With a basic web browser, you've certainly come across various sites where the prospect of capturing, storing and analysing data presented on the pages of the site's pages was considered while browsing the web; Data scientists, whose "raw material" is data, have a lot of possibilities on the web:

- You might want to retrieve an interesting table from a Wikipedia page in order to conduct some statistical analysis.
- Perhaps you're interested in obtaining a collection of movie reviews from a website in order to conduct text data mining, develop a predictive model to detect false reviews, or build a recommendation system.
- You may like to obtain a properties list from a real-estate website in order to create an eye-catching geo-visualization.
- You'd wish to collect additional features to improve dataset using data gleaned from webpages, such as weather data predicting soft drink sales.
- You might be curious on how to conduct a social network analytics on the basis of profile data obtained from a web forum.
- It may be worthwhile to keep an eye on a news site for trending fresh stories on a particular interest topic.

The web contains a plethora of intriguing data sources that serve as a treasure trove of information about a variety of topics. Regrettably, the current unstructured nature of the web makes it difficult to acquire and export this data. While web browsers are excellent at presenting graphics, displaying animations, and arranging websites visually appealing to humans, the web browsers don’t disclose an easy mechanism to export their data, at least not in the majority of circumstances. Instead of accessing webpage through the window of the web browser, wouldn't it be fantastic to gather a rich dataset automatically? Here precisely the site scraping comes into play. If you are familiar with the web, you're undoubtedly wondering: "Isn't this precisely the purpose of Application Programming Interfaces (APIs)?" Indeed, many websites now include an API that enables different user all over the world to access their data sources. For example, LinkedIn, Twitter, Google and Facebook all have different APIs for searching and posting tweets, retrieving a list of your friends and their likes, and seeing who you're connected with. So why do we still require web scraping? The idea is that APIs are an excellent way to access data sources, provided the website in question already has one and the API exposes the capabilities you require. The basic idea is to hunt for an API first and utilize it if possible before embarking on building a web data scraper. For example, rather than reinventing the wheel, you can easily use Twitter's/Facebook API to retrieve data of recent Facebook posts or recent tweets. Nonetheless, there are a variety of reasons that justifies why its preferable to utilize API in web scraping:

- The targeted website for scraping does not provide an API.
- The API does not reveal all of the data you're looking.
- The given API is limited access (limited access time: per day, per seconds, etc.)
- The API is not free.

In each of all that instances, web scraping may be advantageous. The reality is that you can access and retrieve the data via a program if you can view data in your web browser. If data can be accessed via a computer, it may be stored, cleansed, and used in any way desired. (Broucke and Baesens, 2018)

3.4 Application of Web Scraping in Big Data

With the adoption of new technologies, there has been a tremendous rise in the internet users' number and data amount (mainly unstructured) generated by internet users. Due to the fact that scraping is a common method for extracting unstructured data from the Internet, the scraping process is examined when it is subjected to a large amount of data extraction. While scraping enormous amounts of data, we encountered various obstacles, including storage issues, captcha for a huge data amount, the requirement for heavy computation capability, and data extraction dependability. An examination done to a cloud-based web scraping architecture that utilizes Amazon Web Services to manage processing resources and storage with elasticity on demand (DynamoDB and Elastic Compute Cloud). This examination aims to solve both scraping and the feasibility of big data applications with a single cloud-based architecture for data-intensive enterprises. Selenium is mentioned as one of our web scraping tools since it offers web drivers that simulate a real user interacting with a browser. Additionally, assessed performance and scalability of the proposed cloud-based scrapper and discuss the proposed cloud-based scraper benefits over alternative cloud-based scrapers. (Chaulagain *et al*, 2017)

The "big data" term discusses the diverse set of techniques for data collecting and analyzing in methods that were previously unimaginable before contemporary personal computers advent. Due to the automatic collection of information from webpages, web scraping is a method for big data specialists to examine. There can be tens of thousands of variables in a web scraping dataset, but there can also be tens of thousands of examples in a smaller and more manageable dataset in only a few hours using web scraping. The myths surrounding web scraping techniques still debunk that are now being employed to investigate research problems of interest to psychologists. An offered concept called "theory-driven online scraping", that requires the usage of web-based big data be justified by substantial theory. Second, data source theories are defined, a word that refers to the assumptions that a researcher must make about a potential big data source in order to scrape data from it meaningfully. Critically, researchers must develop specific hypotheses to test based on their data source theory, and if these hypotheses are not experimentally supported, plans to use that data source should be modified or abandoned. Thirdly, a case study and sample Python code demonstrating included on how online scraping may be used to collect large amounts of data, as well as links to a web tutorial for psychologists. Fourth, a four-step process outlined for web scraping initiatives. Finally, an exploration done for the legal, practical, and ethical issue explored that arise when performing online scraping operations. (Landers *et al*, 2016).

Due to the massive amount of diverse data generated on a daily basis on the WWW, web scraping is widely recognized as an effective and powerful technique for collecting large amounts of data. To accommodate a variety of scenarios, modern online scraping techniques have evolved from manual, ad hoc operations to the use of completely automated systems capable of converting whole webpages into well-organized data sets. Not only can advanced online scraping technologies parse markup languages or JSON files, but they can also integrate with computer visual analytics and natural language processing to replicate how human users view web information. (Zhao, 2017)

3.5 Application of Web Scraping in cloud computing

Scaling Web Scraping requires clustering numerous Web Scraper instances. While this is tough to accomplish with traditional on-premises servers, it is fairly straightforward with Google and Amazon cloud services as long as the Web Scraping service is packed with Docker. To take advantage of cloud computing's capabilities, (Chaulagain *et al*, 2017) opt to host all resources of scraping on various AWS. In order for web scraping to perform effectively with interactive and dynamic online web pages, it must emulate human actions as closely as possible. In addition, it was discovered that running scraper in parallel is easier when using resources from a scalable cloud platform. Additionally, the resources may be scaled up or down as needed if the load became too heavy. Along with collecting enormous amounts of data in parallel using cloud-based resources, (Chaulagain *et al*, 2017) wish to construct a distributed data processing system using the same resources. The benefit of using a cloud service is that the scraper can work in a way that it can scrape numerous websites concurrently and consume only the resources required. The systems schedule jobs, store retrieved content, and spawn additional resources for the work using Amazon services such as S3 and SQS (Phan, 2019).

A sophisticated technology such as cloud computing enables a web scraper to operate more efficiently. The proposed web scraper which is shown in fig 5 extracts data from the web in a trustworthy manner. Cloud computing enables on-demand access to storage resources and elastic computing in a dispersed environment. The architecture's performance and scalability evaluated using the experiment above on top of Amazon's cloud services. A cloud-based web scraping architecture investigation suggested by Amazon and developed utilizing the EC2, S3, and SQS services for websites scraping using numerous virtual computing machines instances.

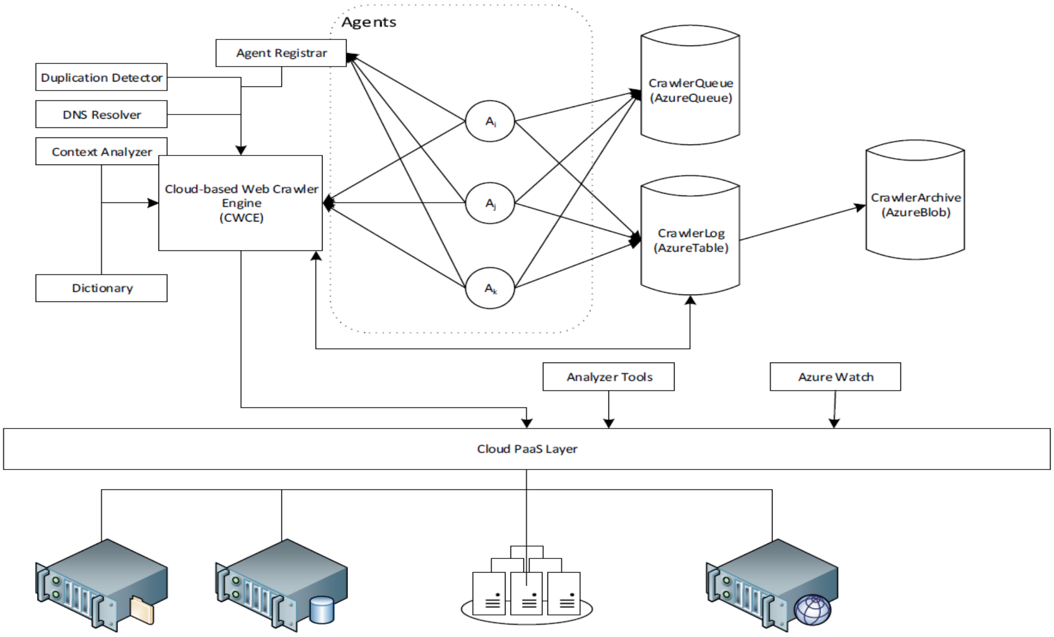


Fig 5: An architecture of the proposed cloud-based web crawler (Bahrami *et al*, 2015)

Additionally, this architecture can be achieved through the use of another cloud service provider. its recommend to use the selenium web renderer as an option if the main goal is improved performance without requiring web automation. Due to the architecture's reliance on entirely scalable resources, it enables the implementation of large data applications. Almost big data solutions resources requirements are included in the suggested architecture.

Web crawlers search the web for interesting and related content on behalf of applications or services. For instance, search engines index the Internet using web crawlers. Web crawlers face various obstacles, including link complexity and extremely intensive computing needs when retrieving complex related links. Another difficulty is the large storage of indexed links and unstructured data such as movies, photographs, and binary files. Due to the continuous growth of data on the Internet and the fact that requests may search material in a format’s diversity including no cloud-based architecture, unstructured data for web crawlers exists in the literature that addresses both storage challenges and highly intensive processing adequately. Cloud computing supports unstructured data and elastic resources, as well as pay-per-use capabilities which enable organizations to create their own limited number of web hosts crawlers or Internet web crawlers. The suggested web crawler enables exploration of the web through the use of distributed agents, with each agent storing its findings on a Cloud Azure Table (NoSQL database). Additionally, the suggested web crawler might use Azure Blob storage to store unstructured and enormous amounts of data. The suggested web crawler's performance and scalability is investigated and discussed its advantages over standard distributed web crawlers.

To obtain an enormous amount of information from the web and maintain unstructured data sets of indexed sites, a web crawler requires extremely expensive computation. Cloud Computing, a high-tech distributed technology, provides the necessary web crawler infrastructure. Environments of Cloud computing provide a distributed architecture which enables applications to consume elastic resources on demand and store huge amounts of data in NoSQL databases with unstructured data support. (Bahrami *et al* ,2015).

Some details about web scraping and cloud computing partially discussed in the section 3.4.

3.6 Application of Web Scraping in Cyber Security

Semantic Web technology can be applied to the subject of cybersecurity. Cybersecurity is certainly a very complicated and expansive topic, with actions divided into two categories: those conducted to design an optimal system with the fewest feasible vulnerabilities and those undertaken in response to problems that arise once the system is operational. While the first type of activity takes a fairly standard method to enhancing security for programming writers, the second is a cat and mouse game in which, after a black hat hacker discovers (and exploits) a type of weakness, system experts try to resolve it. In comparison to white hat hackers, black hat hackers obtain illicit access to and perform acts on a computer system without the owner's authorization in order to achieve personal benefits. One of the articles' purposes is to identify and discuss the actions that can be taken between the two types of activities outlined above, using Semantic Web technology. (Georgescu and Smeureanu, 2017)

Additionally, the CFAA has produced frictions due to its interaction with a site's terms of service (ToS). This conflict is around the question of whether authorization is canceled based on statements included in these lengthy, legalistic documents which is rarely read by users. For example, such papers frequently prohibit online scraping in broad, imprecise wording (Fiesler *et al*, 2020) and despite over sixty judicial views over the last two decades, the legal position of scraping remains "just shy of unknowable, or fully left to the whims of courts" (Sellars, 2018). This creates significant difficulties for businesses, researchers, and journalists, as computer misuse law has the potential to effectively convert prospective civil accountability for break of agreement into criminal accountability under the CFAA (Veale and Brown, 2020).

4 Web Scraping and Legal concerns

Automatic data extraction (Web Scraping) is becoming more prevalent in corporate and academic research projects. Web Scraping has been facilitated through the development of a variety of tools and technologies. Regrettably, the legal and ethical implications of employing these tools for data collection are frequently disregarded. Inadequate consideration of these Web Scraping factors can result in major ethical disputes and lawsuits. This work conducts a review of the legal literature, as well as the literature on ethics and privacy, in order to highlight general areas of concern, as well as a list of particular problems that scholars and practitioners engaged in Web Scraping should address. Reflecting on these issues and concerns may assist researchers in reducing the risk of ethical and legal conflicts arising from their work. (Krotov and Silva, 2018)

The legal landscape surrounding web crawling and scraping is still developing, and courts are only beginning to address claims arising from web scraping or crawling for analytics reasons. Furthermore, determining whether crawling or scraping for the purpose of analytics raises legal problems is a highly fact-specific determination. However, the incidents to date, including the two listed above, reveal a number of difficulties which website owners and those performing analytics utilizing data acquired from web-based sources should address, including the following:

a. the language used in the service agreement or terms of use, as well as whether such terms declare automated access to the website, the use of any data gathered through such means, and the use of the website for purposes other than the non-commercial use, user's personal;

b. the terms of use enforceability, for example, whether they are displayed to the user via a clickwrap mechanism requiring the user to agree or not to those terms, or via a terms of use page accessible via a prominent link on each other page of the website and indicating that any the website use is conditional on the user agreeing to those terms;

c. the use of technological tools to prevent unauthorized scraping or crawling or to establish crawl rates, such as the robots.txt protocol;

d. whether the website access is safeguarded in a way that allows for an alleged violation of the CFAA or California's Penal Code Section 502;

e. whether the data of the website's content is copyright protected; and

f. whether the owner of the website intends to permit or license content usage.

It is unavoidable that scraping and crawling for analytics purposes will be continuously evolving, and that courts will wrestle with the legal theories and facts applicable to scraping and crawling situations. While the law in this area remains to evolve, both scrapers and website owners should be aware of precedent-setting decisions and careful about remaining current on potential developments. (Snell and Menaldo, 2016)

5 The Future of Web Scraping

As computers become more and more powerful on average, the potential for more complex and intelligent web scraping algorithms continues to rise. (Asma, 2020)

One important aspect will be the increased usage of artificial intelligence and machine learning in web scraping techniques. At the present, most web scrapers used straightforward methods which work on many, but not all, target sites. Utilizing machine learning could result in scraping techniques which automatically "learn" from experience and can adapt to a wider variety of sites. (Suganya, 2019)

One major issue which will need to be addressed very soon is the legal standing of web scraping. As with so many internet-based technologies, it's hard to define what is and isn't legal based on laws that predate the internet. As we move forward into the 21st century, legal bodies around the world have continued to adapt themselves to the information age, and soon a consensus will have to be reached on whether any or all instances of web scraping violate the right to information privacy. (Krotov and Tennyson, 2018)

In the event that web scraping continues to grow as an industry, it will be important for any business which has not caught on to join in soon, both in regards to protecting their own data they don't want accessed, and in investing into their own web scraping pursuits. Companies which fail to live up to the modern standards expected will often become obsolete themselves and be rejected by consumers. (Asma, 2020)

6 Results and Discussion

In web scraping, code reuse and maintenance are especially critical. Code reuse is important in web scrapping because when we develop program to mostly have some common feature that we used before For example for website data scraping, the program need to access the website so if you built a good code to access a site using username and password, you can reuse it in a future project if you are going to run it in the same situation. Web scraping involves code maintenance because, no matter how reliably you construct your scraper, a significant change in the targeted site's design or behavior might make it extremely difficult for your program to find the appropriate data on the site. If you have a web scraping procedure that runs on a regular basis for any length of time, you'll need to be ready to make code changes to compensate for changes on the destination web site. Finally, Python is perfect for easy implementation. Web scraping is frequently required in order to go to more interesting aspects of a project. Simple web scraping in Python can be pulled up fast, especially if you have a large library of reusable code. (CreateSpace Independent Publishing Platform, 2015)

It was discovered that the web scraping has a wide variety of applications across the field of Business Intelligence. Using web scraping programs would involve a higher upfront cost, but would pay for themselves over and over again as a form of short-term loss, long term gain. Furthermore, using a web scraping program would yield data that is far more thorough, accurate, and consistent than manual entry. However, it is important to use it in tandem with other data processing tools, as the raw data on its own will have far too many extraneous details to be useful unless properly filtered and processed.

Although online scraping is an effective method for obtaining massive data sets, it has some ethical and legal consideration, scrapping may happen over copyrighted data and lead to copyright infringement furthermore aggressive spidering usually leads to large number of requests to sites that may cause loss of services which is naturally undesirable by site owners. The use of robots.txt allows some form of self-compliance in this method website owners place a text file in their root folder that lists all the files they forbid from being crawled. Compliance isn’t necessarily enforced in these cases but ethical scrappers should abide by site owner wishes by abiding with robots.txt (Krotov and Silva, 2018)

Proper cleaning and preprocessing are a requirement for most web scrapped date, performing this process required manual study of the original data format and structure (Tarannum, 2019).

There have been safety protocols implemented that use machine learning to fend off against malicious attacks and hackers. (Zhao, 2017)

Based on the literature reviewed web scrapping is a valuable tool that can allow us to extract meaningful knowledge from otherwise unstructured data. Combining web scrapping with other technologies such as big data can yield further value when performing data extraction in scale. (Chaulagain *et al*, 2017).

However, there has been controversies surrounding web scraping as users have plagiarized whole websites and claimed the data as their own, resulting in lawsuits and copyright infringements. To avoid any further conflicts in the future, any company or organization that engage in web scraping must be under surveillance or be subjected to some kind of process that can detect plagiarism.

Dealing with embedded tuples ensures that individual records in datasets are secure. (Tarannum, 2019)

Providing a web API such as REST or SOAP is a better approach than web scrapping as it allows users to receive structured data in JSON or XML format without the need of extraction from text, it also allows us to avoid spidering and large number of requests usually associated with bot scrapping (Alrashed *et al*, 2020).

It can be suggested that cloud computing can provide a better facility for web scraping tools as it is more reliable and the user can access it at any given time when it is needed. (Chaulagain *et al*, 2017)

Natural Language Processing (NLP) is another way web scrapping can be improved with libraries included in many computer languages that can be used for data cleaning, filtering, organizing such as Python and R languages. (Krotov and Silva, 2018)

Using web scraping is highly recommended as it is cheaper as compared to other methods of data collection given that it is done in an ethical manner and no one is harmed. (Hillen, 2019).

The main difficulties that need to be examined before apply web scraping:

- Are the data collected from human subjects? If yes, is data scraping ethical?
- Is there an API available on the website?
- Is web scraping permitted on the website?
- Is the data presented in the form of an HTML table?
- Is it simple to select CSS selectors using the Selector Gadget?
- Is there any data that is not numerical? If yes, how manipulee is it?
- Would the scraping procedure entail iteration across numerous pages? how much data do you intend to scrape if that is true? just a sample or the entire site?

7 Conclusions

Web scraping is a highly useful tool in the information age, and an essential one in different fields for any company wishing to maintain their online presence, which in of itself is very necessary to survive in today’s market. While more strict legal measures may be enforced in the coming years, the rate of this new market will keep continuing to grow, and it is therefore an incredibly valuable skill to learn. One of the best languages for implementing it in would be Python, due to its fast learning curve and powerful syntax.

ACKNOWLEDGEMENTS

The author is grateful to the Applied Science University, Bahrain for the full financial support granted to this research paper.

References

[1] Almaqbali, I. S., Al Khufairi, F. M., Khan, M. S., Bhat, A. Z., Ahmed, I. (2019). Web Scrapping: Data Extraction from Websites. Journal of Student Research.

[2] Alrashed, T., Almahmoud, J., Zhang, A. X., Karger, D. R. (2020). ScrAPIr: Making Web Data APIs Accessible to End Users. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–12). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3313831.3376691

[3] Asma. (2021, September 29). The Future of Web Scraping Services. Retrieved from Information Transformation Services: https://it-s.com/the-future-of-web-scraping-services/

[4] Banerjee, R. (2014). Website Scraping. Happiest Minds Technologies Pvt. Ltd.

[5] Bhatia, M. A. (2016). Artificial Intelligence–Making an Intelligent personal assistant. Indian J. Comput. Sci. Eng, 6, 208-214.

[6] Broucke, S. V., Baesens, B. (2018). Practical Web Scraping for Data Science: Best Practices and Examples with Python. (1st, Ed.) Apress.

[7] Chamoso, P., Bartolomé, Á., García-Retuerta, D., Prieto, J., De La Prieta, F. (2020). Profile generation system using artificial intelligence for information recovery and analysis. Journal of Ambient Intelligence and Humanized Computing, 11(11), 4583-4592.

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

[9] CreateSpace Independent Publishing Platform. (2015). Learn Web Scraping with Python in a Day: The Ultimate Crash Course to Learning the Basics of Web Scraping with Python in No Time. CreateSpace Independent Publishing Platform, North Charleston, SC, USA.

[10] Dogucu, M., Çetinkaya-Rundel, M. (2021). Web scraping in the statistics and data science curriculum: Challenges and opportunities. Journal of Statistics and Data Science Education, 29(sup1), S112-S122.

[11] E. Suganya, S. V. (n.d.). Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms. International Conference on Intelligent Systems Design and Applications. 2019.

[12] Farley, E.J. and Pierotte, L. "An Emerging Data Collection Method for Criminal Justice Researchers." Justice Research and Statistics Association, December 2017

[13] Fiesler, C., Beard, N., Keegan, B. C. (2020). No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. Proceedings of the International AAAI Conference on Web and Social Media, 14(1), 187–196.

[14] Georgescu, T. M., Smeureanu, I. (2017). Using Ontologies in Cybersecurity Field. Informatica Economica, 21(3).

[15] Gheorghe, M., Mihai, F.-C., Dârdală, M. (2018). Modern techniques of web scraping for data scientists. International Journal of User-System Interaction, 11, 63–75.

[16] Grasso, G., Furche, T., Schallhart, C. (2013). Effective Web Scraping with OXPath. Proceedings of the 22nd International Conference on World Wide Web (pp. 23–26). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2487788.2487796

[17] Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., Perez-Meana, H. (2018). A web scraping methodology for bypassing twitter API restrictions. arXiv preprint arXiv:1803.09875.

[18] Hillen, J. (2019, November). Web scraping for food price research. British Food Journal, 121, 3350–3361.

[19] Kasereka, H. (2020). Importance of web scraping in e-commerce and e-marketing. SSRN Electronic Journal.

[20] Krotov, V., and Tennyson, M. 2018. “Scraping Financial Data from the Web Using the R Language,” Journal of Emerging Technologies in Accounting, Forthcoming.

[21] Krotov, V., Silva, L., 2018. Legality and ethics of web scraping, Twenty-fourth Americas Conference on Information Systems, New Orleans..

[22] Landers, R. N., Brusso, R. C., Cavanaugh, K. J., Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological methods, 21(4), 475.

[23] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.

[24] M. Bahrami, M. Singhal and Z. Zhuang, "A cloud-based web crawler architecture," 2015 18th International Conference on Intelligence in Next Generation Networks, 2015, pp. 216-223.

[25] Manjushree, B.S and Sharvani, G.S, 2020. Survey on Web scraping technology. Wutan Huatan Jisuan Jishu, XVI(VI), pp.1-8.

[26] Milev, P. (2017). Conceptual approach for development of web scraping application for tracking information. Economic Alternatives, (3), 475-485.

[27] Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web. O'Reilly Media, Inc.

[28] Parikh, K., Singh, D., Yadav, D., Rathod, M. (2018). Detection of web scraping using machine learning. Open access international journal of Science and Engineering, 114–118.

[29] Persson, E. (2019). Evaluating tools and techniques for web scraping.

[30] Phan, H. (2019). Building Application Powered by Web Scraping.

[31] Poojitha Thota and Elmasri Ramez. 2021. Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis. In The 14th Pervasive Technologies Related to Assistive Environments Conference

(PETRA 2021). Association for Computing Machinery, New York, NY, USA, 306–314.

[32] Saurkar, A. V., Pathare, K. G., Gode, S. A. (2018). An overview on web scraping techniques and tools. International Journal on Future Revolution in Computer Science & Communication Engineering, 4(4), 363-367.

[33] Scraping Expert. (2021, September 30). Advantages and Disadvantages of Web Scraping. Retrieved from Scraping Expert: https://scrapingexpert.com/advantages-disadvantages-web-scraping/

[34] Sellars, A. (2018). Twenty Years of Web Scraping and the Computer Fraud and Abuse Act. Boston University Journal of Science & Technology Law, 24(2), 372. https://scholarship.law.bu.edu/faculty_scholarship/465/

[35] Sirisuriya, D. S., (2015). A comparative study on web scraping. Proceedings of 8th International Research Conference, KDU.

[36] Snell, J., Menaldo, N. (2016). Web scraping in an era of big data 2.0. Bloomberg Law News.

[37] Tarannum, T. (2019). Cleaning of web scraped data with Python. Ph.D. dissertation.

[38] Uzun, E. a. (2018). Fundamental Sciences and Applications. 87.

[39] Veale, M., Brown, I. (2020). Cybersecurity. Internet Policy Review, 9(4), 1-22.

[40] Vording, R. (2021). Harvesting unstructured data in heterogenous business environments; exploring modern web scraping technologies.

[41] Yannikos, Y., Heeger, J., Brockmeyer, M. (2019). An Analysis Framework for Product Prices and Supplies in Darknet Marketplaces. Proceedings of the 14th International Conference on Availability, Reliability and Security. New York, NY, USA: Association for Computing Machinery.

[42] Zhao, B. (2017). Web scraping. Encyclopedia of big data, 1-3.

Notes on contributors

Moaiad Ahmad Khder is (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Information Science and Technology, The National University of Malaysia, in 2015. He is currently an Assistant Professor with the Computer Science Department, Applied Science University, Bahrain. He has been working on the area of mobile environment, mobile database, data science, big data and cloud computing.