International Journal of Hybrid Information Technology
Vol.2, No.1 (2022), pp.47-64
http://dx.doi.org/10.21742/ijhit.2653-309X.2022.2.1.04

A Web Scraping Framework for Descriptive Analysis of Meteorological Big Data for Decision-Making Purposes

Abderrahim El Mhouti[1*], Mohamed Fahim[2], Adil Soufi[3] and Imane El Alama[4]

[1,2]ISISA, FS, Abdelmalek Essaadi University, Tetouan, Morocco
[3,4]FSTH, Abdelmalek Essaadi University, Tetouan, Morocco
[1]a.elmhouti@uae.ac.ma, [2]m.fahim @uae.ac.ma, [3]a.soufi@uae.ac.ma,
[4]i.elalama@uae.ac.ma

Abstract

With the large increase in the amount of heterogeneous, complex, and unstructured data issued from Web sources, there is an emerging need to develop Big Data technologies and tools to extract and manage this data. In this context, Web scraping for Big Data is a technique that has gained importance because of its rapidity and efficiency in gathering data for Big Data technology. This study was conducted to propose a Web Scraping framework for descriptive analysis of meteorological Big Data. The introduced framework makes it possible to extract a set of data to process it, present it in a form that is easier to analyze and understand, and use it for decision-making purposes about weather forecasts. The study employed descriptive analysis of available big data as it allows one to easily make quality and effective decisions. The web scraping process takes place in several stages, including data extraction, data archiving in a data warehouse, and finally data filtering and analysis. To test its applicability, the proposed web scraping framework was implemented and tested in the meteorological context to extract and present meteorological data issued from a specialized web source. The proposed system makes it possible to restore the data in the form of statistical models published in a dashboard. The results of the study revealed that the predictive models provided by the system are capable of predicting certain weather-related variables, such as humidity, precipitation, and temperature. The opportunities and implications to leverage the results of this study are many, including weather forecasting and decision support.

Keywords: *Analysis, Extract, Load, Meteorological big data, Web scraping, Transform*

1. Introduction

Today, the emerging evolution of the Internet has made the web one of the largest sources of public data in the world. Every day, several petabytes of information and data are published by Internet users and managed via the Internet in various formats, such as HTML (HyperText Markup Language), PDF (Portable Document Format), CSV (Comma Separated Values), or XML (Extensible Markup Language) document [1]. At this rate of technological evolution, more and more data is being generated from various web sources. When we talk about such a quantity of generated data, we are faced with a concept known as "Big Data" which is described by the concept of the "3Vs": Volume, Variety, and Velocity [2].

Article history:
Received (August 5, 2022), Review Result (September 2, 2022), Accepted (October 15, 2022)

Nevertheless, this large source of information represents a huge challenge for organizations and researchers involved in data mining and data analysis issues [1][3]. With a huge amount of sparse, unstructured, and heterogeneous data, researchers are faced with many problems related to the processes of collecting, managing, and analyzing this information in a reasonable time to use it for prediction and decision support.

In this context, "Big data", as an abstract concept [4], arose to describe this huge amount of generated data in the world generally varied and unstructured. Around the world, the use of big data by researchers and policymakers has become increasingly widespread [2]. Over the last few years, big data is widely used in both business and public policy decisions [5]. In addition, many institutions around the world use big data statistics in their policies in various fields such as transportation [6], health [7], education [8], agriculture [9], etc. For its part, in recent years, Morocco has launched many initiatives to exploit the opportunities of big data. However, despite the efforts made, big data in Morocco is still in its initial state [10][11]. So, the rapid emergence of the concept of big data around the world has led to the emergence of many big data techniques and tools, allowing one to collect information and convert it into structured data. One of the main technologies used to access the data is the web scraping technique, which represents one approach to big data with great potential. Web scraping, which is turning into a great data access technique these days [12][13], focuses on the transformation of unstructured data issued from the web into structured data that can be stored and analyzed in a central local database or spreadsheet [14].

Indeed, web scraping is an approach to big data that has great potential in the automated collection of information from web sources. While big data scraping can create huge datasets, it also has valuable potential for data analysis and visualization to support predictive and decision-making processes. Several research works using web scraping for decision support have been realized in this context. We mention among others, the support of urban design decision processes [15], support for tourism decision-making [16], support for adaptation to climate change, [17], etc.

Thus, as web scraping is one of the major sources for the extraction of unstructured data from the Web [18], this paper is particularly interested in the web scraping process in a big data context. The aim is to allow the extraction of a large mass of meteorological data from web sources, to be able to manage them, present them in a simpler form to analyze and understand them and then use them for prediction and decision support. The use of this weather data for prediction and decision support can range from simple day-to-day decisions about various needs (short term) to large-scale emergency response (medium and long term). For example, in the short term, this data is widely used as a decision-support tool. Beneficiaries can be people who want to know what clothes to wear or who plan their weekend according to weather conditions (temperature, sunshine, or precipitation), people involved in weather-dependent socio-economic activities, farmers who pay close attention to weather forecasts in planning their daily tasks (e.g., they carefully monitor fallen and forecasted precipitation before deciding to water), etc. For medium- and long-term planning, these statistical and dynamic data can be used to characterize the risks of hazardous events and to make decisions regarding vigilance.

In line with this, the study was conducted to propose setting up a web scraping framework for descriptive analysis of meteorological Big Data. The aim o the study is to offer not only meteorological data analysis but also decision support related to weather forecasts and also various prediction capabilities (of certain weather-related variables, such as humidity, temperature, and precipitation) as described above. The proposed framework makes it possible to restore the data in the form of statistical models published in a dashboard. The web scraping process takes place in several stages, including data extraction, data archiving in a data warehouse, and finally data presentation and description. To test its feasibility, the proposed framework is implemented as a web scraping system which was used to extract meteorological data and present it in a simpler form for decision support purposes.

The remainder of this paper is structured as follows: the second section deals with related works covering some of the recent studies relating to big data concepts and web scraping tools and issues. The third section looks at the methodological approach adopted for the design and modeling of the web-scarping framework of descriptive analysis of meteorological Big Data. This section states also the web scraping framework implementation and its application in the meteorological context. The next two sections present and discuss the results obtained. The last section concludes the study and outlines future works.

2. Literature review

Today, in various fields, obtaining useful big data effectively from the Internet and making the most out of it are essential in decision support. However, with over 2 billion Web pages on the Internet, manually collecting big data is not feasible. In this sense, the Web scraping technique has gained importance because of its efficiency in gathering big data. Web scraping for big data is an approach that has great potential in collecting varied information from web sources. This potential is due to the ability of this technique to automate the data extraction process, its speediness, and its low cost for obtaining accurate, clean, and structured data.

One of the main areas of application of web scraping for big data is data analysis and visualization to support prediction and decision-making processes. In this context, several research works using big data scraping for decision-support purposes have been carried out.

In this section, and after the presentation of the two key concepts on which this work is based (big data and web scraping), a literature review study focusing on the use of web scraping technologies for big data extraction for decision support purposes will be explored.

2.1 Big data concept

2.1.1. Overview

The term “Big Data” is introduced in 2005 by Roger Magoulas to define a great amount of data available that cannot be manipulated by traditional data management techniques due to the size and complexity of this data [19]. The big data paradigm refers to large-scale data tools and infrastructures addressing new requirements in processing data volume, velocity, and variability. If traditional databases systems are designed to address smaller volumes of structured data, with consistent data structure and predictable updates, the big data concept is interested in the variety of diverse formats of data with both batch and stream processing in several areas such as structured data, geospatial data, audio, and video data, 3D data, unstructured text including sensor data, log files and social media [20]. Thus, big data is defined by its size, comprising a large, complex, and independent collection of data sets, and where this data cannot be managed with standard data management techniques.

2.1.2. Concepts and characteristics

In addition to the size of the data, the concept of Big Data also includes data variety and speed. These three properties together (volume, velocity, and variety) form the dimensions of Big Data (or three V of big data). However, as shown in [Figure 1], most studies expand this concept to five key characteristics (5 Vs), namely, *volume*, *velocity*, *variety*, value, and *veracity* [21].

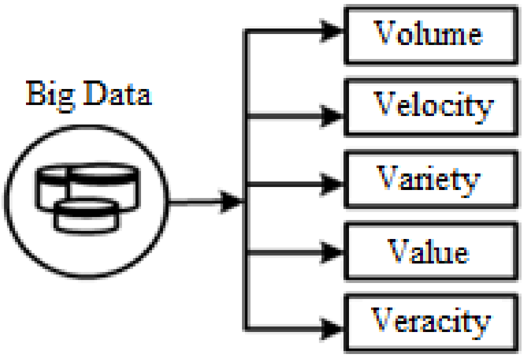


Figure 1. Big data concepts [21]

- Volume: refers to the quantity of data gathered by an organization. Data can come from every possible source: networks, sensors, engines, etc. The challenge is that is very hard to sustain, store, analyze, and ultimately use these data.
- Velocity: concerns the time during which the information can be processed. The challenge is that the resources for analyzing data are limited compared to the volume of data, but the requests for information are unlimited because some activities are very important and need immediate responses.
- Variety: this 3rd dimension refers to the type of data that big data can store because there are many types of sources, and the data that came from them is varying very much in size and type. This data can be structured or unstructured.
- In addition to the three Vs above, big data is characterized by other important properties:
- Value: the fourth "V" value, which for some researchers represents the most vital and irreplaceable characteristic of all the "Vs" in Big Data, as they believe it has the power to turn industry data into valuable information.
- Veracity: this 5th dimension refers to the degree to which a leader trusts the user data to take a decision. Indeed, veracity is a difficult task to achieve with big data, because due to the volume of data and its variety is very difficult to identify accurate and useful data from the "dirty data". The challenge here is that the "dirty data" can lead to incorrect results and/or different types of errors, which can affect the velocity dimension of big data [22].

2.1.3. Infrastructure and data collection and extraction techniques

From a practical perspective, research efforts are being made to establish comprehensive standards and certifications for big data technology. However, many big data systems have emerged in recent years. Thus, several systems, such as Hadoop [23], Spark [24], and Flink [25] are designed and developed to be general-purpose big data systems, while several others, such as GraphLab [26] and SciDB [27] are implemented to serve specific types of workloads.

On the other hand, there are many data collection and extraction techniques used in big data technology to accumulate data. One of the major sources of big data is collected directly from web environments (sites, platforms, web applications, etc.) using the web scraping method. This technique consists of extracting and storing data from a site in a structured way and thus makes it possible to reuse this data.

Web scraping is an automated process that any organization can use to efficiently collect large amounts of targeted data from different web sources. Thus, the concept of big data is closely related to the concept of web scraping. Web scraping is one of the reliable data collection techniques for big data technology.

In the following, the paper presents a brief review of web scraping for use in the collection of big data.

2.2. Web Scraping for big data

2.2.1. Overview

Web scraping is a process of collecting automatically information from Web sources. It is a domain with active developments that share a common goal with the vision of the Semantic Web, an innovative idea that still requires breakthroughs in data processing, artificial intelligence, semantic understanding, and human-computer interactions.

Furthermore, web scraping is one of the techniques to consider if we want to work in the field of big data. Big data scraping is a process for crawling the web and collecting target data on a large scale from different web sources (sites, platforms, pages, etc.). It automates the data collection process and the conversion of the scraped data into different formats, such as HTML, CSV, Excel, JSON, text, etc. [Figure 2] below illustrates the Web Scraping process.

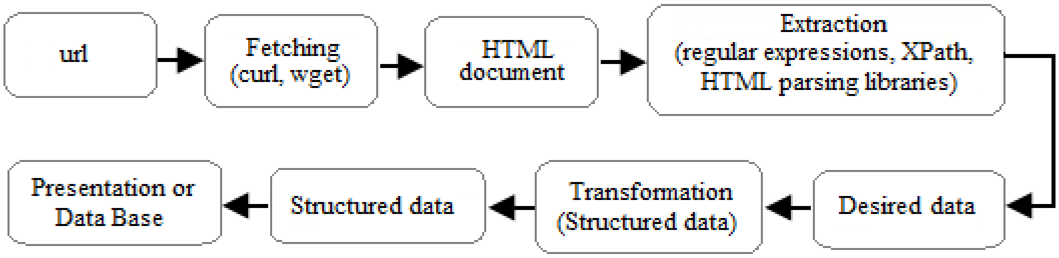

url
Fetching (curl, wget)
HTML document
Extraction (regular expressions, XPath, HTML parsing libraries)
Desired data
Transformation (Structured data)
Structured data
Presentation or Data Base

Figure 2. Web scraping process [28]

The scraping process generally simulates web crawling by integrating appropriate web browsers or implementing a low-level hypertext transfer protocol. Also, web scraping is closely related to the web indexing technique which consists of retrieving the web information and is adopted by several search engines to index information on the web through a bot. Web scraping for big data, on the other hand, focuses on crawling the web, collecting large-scale data, and transforming unstructured data on the web, usually HTML, into structured data that can be stored and analyzed in a central local database or spreadsheet [29].

The web scraping process involves a set of operations that are commonly known as ETL (Extract, Transform, Load). ETL methods extract information from various business processes and feed it into a data warehouse or a database. The ETL process involves the following phases:

- Extraction: consists of extracting data from the source data stores or even data coming in a streaming fashion. Typically, only the data that are different from the previous execution of an ETL process should be extracted from the sources.
- Transformation: once the extraction is done, data accumulated are propagated to a special-purpose area of the warehouse, called the Data Staging Area (DSA), where these data are transformed, homogenized, and cleaned.
- Loading: the data transformed are loaded to the central data warehouse and all its counterparts. In the data warehouse setting, the ETL process refreshes the data warehouse periodically.

Finally, the data is analyzed and presented in various forms such as graphs. The goal is to use this data for different purposes, including prediction and decision-making purposes.

2.2.1. Extraction modes and techniques

We can make a basic distinction about web scraping according to data extraction mode:

- Manual extraction: when browsing the web, the user can extract information relevant to his interests from the web pages visited. The most common practice for this type of data extraction is simple copy/paste.
- Semi-automatic extraction: the user can aspire/clean the elements of one or more web pages corresponding to his interests using software or a web application.
- Automatic extraction: the extraction process is done completely automatically. This is thanks to the emulation, by a machine, of a web browser, which visits the pages and which makes it possible to follow the various links to automatically generate a corpus of linked pages. This mode is essential for gathering big data sets.

In the context of automatic or semi-automatic extractions modes, what is essential is to identify in the documents analyzed the data of interest to separate them from all the content. Furthermore, web scraping for big data requires more advanced approaches and techniques. Here is a non-exhaustive list of techniques that can be used to extract the data [30]:

- Regular expressions: there are features available in virtually any programming language that can identify patterns within the textual content. Thanks to a syntax that allows combining several analysis rules at the same time, it is possible to extract elements that correspond to the criteria defined in the regular expression.
- XPath: it is a W3C standard used to find elements in an XML document. This language exploits the hierarchical structure of the nodes (and attributes) of an XML document and therefore requires a very precise document structure.
- DOM: this technique is similar to XPath because it also exploits the hierarchical structure of a web page across the DOM. Once again, this is a W3C standard that provides access to the content of the different tags of an HTML page thanks to their hierarchical positioning on the page.

2.2.3. Web scraping tools

Many software tools and techniques for doing web scraping are available or are custom designed for users to extract desired information required from millions of websites. The most common technologies used for scraping are Selenium, cURL, Wget, HTTrack, Scrapy, Node.js, and PhantomJS. Of course, these are just a few, and there are hundreds of Web scraping tools and services for anyone who wants to illegally scrape data from popular websites.

[Table 1] presents a list of "low-level" Web scraping tools, which integrate scraping into larger programming logic.

Table 1. Web scraping tools

Scraping tool	Language	Description
Rvest	R	R package that makes it easy to import data from web pages.
Goutte	PHP	PHP Library for Web scraping/crawling.
jQuery	JavaScript	Library that allows, thanks to asynchronous requests (AJAX), to do scraping on the client side.
Scrapy	Python	Python Library for web scraping/crawling.
Selenium WebDriver	Java …	API to program actions on the interface, check the answers, also scrapper the data.

To facilitate the process of further processing the collected data, most scraping software tools are written in Python language using frameworks and libraries for webs crawling, such as Scrapy, Ghost, lxml, aiohttp, or Selenium.

2.3. Web Scraping for big data: Literature review

The web scraping technique is used in different activities, including big data purposes. Indeed, web scraping for big data is an approach that has great potential in the automated collection of varied information from web sources. One field of application of this approach (web scraping combined with big data) is data analysis and visualization to support prediction and decision-making processes. In this context, several research works using big data scraping for decision-support purposes have been carried out.

Ensari and Kobas [15] propose methods to collect and visualize urban data to support urban design decisions. The study is based on web scraping techniques to collect a variety of publicly available data in the municipal boundaries of Kadıköy in Istanbul as well as visual programming tools to map and visualize this information. By overlaying the resulting maps, the authors can visually communicate urban conditions, including demographic and economic trends based on online real estate listings as well as the spatial distribution and accessibility of public and commercial resources. The proposed approach has valuable potential to support urban design decision-making processes.

Adhinugroho et al [16] investigate the use of big data, through web scraping, to produce tourism statistics, specifically accommodation statistics in Indonesia using online travel agent data. Using an accommodation directory compiled based on the collected big data, the authors were able to determine room occupancy rates and a weighted average of daily and monthly accommodation prices. This data can be used for future decision-making.

Ford et al [17] examine how big data, using information scraped from the Internet, can inform adaptation research and decision-making and describe what the adaptation community needs to maximize this opportunity. The authors argue that the careful application of big data could revolutionize our understanding of how to manage climate change risks.

Other research work focuses on the analysis of massive business transaction data, obtained via web scraping, for better decision-making [28]. In this context, market analysts and business intelligence professionals rely on high-quality and meaningful data to accomplish their tasks. This makes web scraping a feasible approach for business operations, including market pricing, market trend analysis, entry point optimization, and competition monitoring [31].

In addition to prediction and decision support, web scraping for big data has interested many institutions and researchers in different fields. For example, in the academic research field, the web is a data resource that concerns several domains and a large amount of data available sometimes makes it possible to fill the methodological biases related to the information available on the web (truth, authenticity, etc.). In this context, a model for both scraping and feasibility for big data applications in a single cloud-based architecture for data-based industries was proposed by Chaulagain et al [18]. Other research works focus on the evaluation of web scraping tools [32] and the extraction of Big Data from the Internet for use in psychological research [33].

In the commercial field, big data scraping is used to compare the information available between competing sites (e.g. airline tickets of different companies for the same journey). In this context, a framework of crawling and scraping methods on an e-commerce website to obtain HTML data was introduced by Onyenwe et al [34] to identify product updates based on the current time. Also, a study is carried out by Kasereka [35] to explore the importance of web scraping in e-commerce and e-marketing. The authors explain the advantage of the web scraping technique and provide a practical example that can be beneficial for e-commerce businesses and online marketers.

In the marketing analysis, the "traces" left on the internet are more and more numerous and provide an idea of our preferences, habits, etc. These data can very well be exploited for marketing analysis (especially in the context of online advertising). In a study conducted by Herrman and Hoyden [36] where the authors state that web scraping has become an essential application in marketing and data science. In addition, the authors highlight the importance of open data and social media data as a target for scraping and illustrate examples of open data and social media integration, sentiment analysis, and classification of content from websites as a use of web scraping in a market research environment.

As for Morocco, it has launched many initiatives to exploit the opportunities of big data in recent years. However, despite these efforts, big data in Morocco remains basic and still in its infancy [10][11]. In 2021, Morocco is thinking big for data centers. In this sense, the Mohammed VI Polytechnic University of Benguerir proceeded in February 2021 to the inauguration of its new Data Center housing the most powerful "supercomputing" in Africa (African Supercomputing Center). Located in the heart of the green city of Benguerir and spread over an area of 2,000 m2, this Data Center will increase the capacity for scientific experimentation and thus allow for better control of the big data collected in Morocco [10].

3. Proposed method

Big data scraping is widely used for data analysis purposes. In this sense, this study consists in proposing a Web Scraping framework for descriptive analysis of meteorological Big Data issued from web sources and used for prediction and decision support purposes.

In this section, this paper details the methodology adopted to allow other researchers to replicate the study. Thus, the paper presents the conceptual and functional study of the proposed Web Scraping approach for descriptive analysis of meteorological Big Data and addresses framework architecture. This work pays particular attention to the capture of the framework's functional needs. The paper also describes the technical needs and the implementation of the Web scraping framework to allow its experimentation thereafter.

We believe that this methodology is appropriate for this research because it will contribute to the implementation of a weather data analysis system that will be a new possibility for users who wish to base their decisions on several forecast sources. On the other hand, the methodology followed will allow other researchers to easily replicate the study in other contexts.

3.1. Framework architecture

In this work, the architecture of the proposed framework is based on a set of modules ensuring big data scraping by collecting data from websites. The framework architecture uses the ETL process based on the Talend tool to ensure data processing. Web scraping, provided by the Web scraping service module, is done in three main steps: useful data extraction, data transformation, and data loading. Data collected is stored in a data warehouse. [Figure 3] shows the proposed web scraping framework architecture.

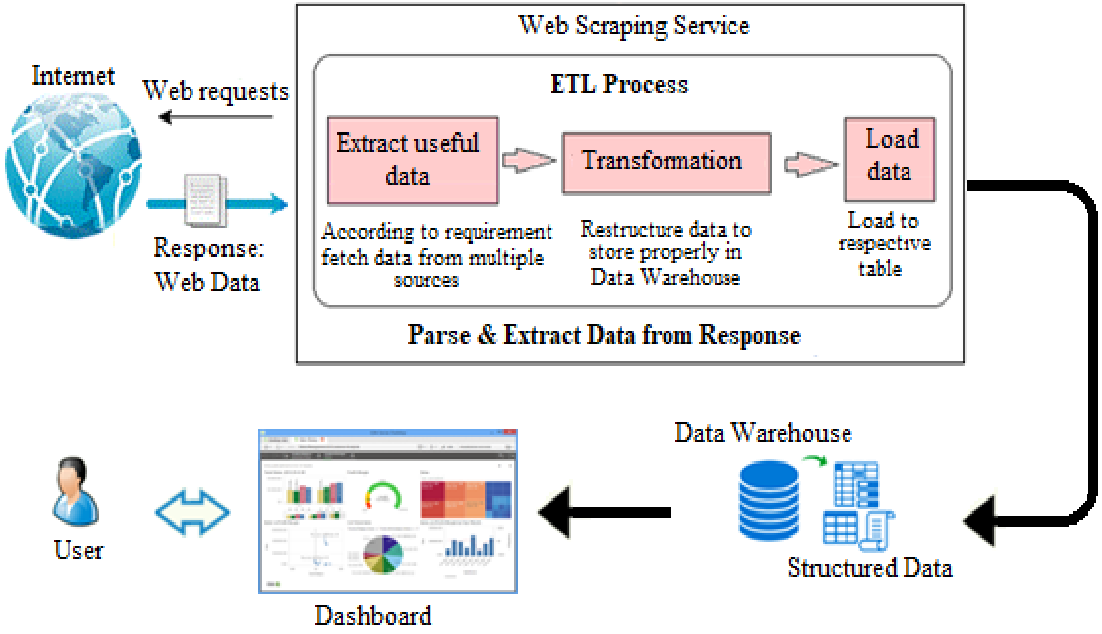


Figure 3. Web scraping system architecture

In this work, the main actor in interaction with the proposed web scraping framework is the Dashboard user. The Dashboard user can perform multiple tasks. It can view the different indicators and filter the data under several axes. Also, the Dashboard user can view the key performance indicators and can filter data under several analysis axes. It can also cancel the filter or download reports.

Data stored in the data warehouse is presented as Dashboard using the Qlik Sense tool. Qlik Sense is a comprehensive data analytics solution that sets the benchmark for a new generation of analytics. It is based on a unique associative analysis engine and an extremely powerful cloud platform. The Qlik Sense interface helps to understand data, use it effectively and make better decisions.

3.2. Database modeling

To store meteorological data, the proposed web scraping framework uses a Data Warehouse. The latter is the intermediate data storage site for the constitution of the decision-making information system. The Data Warehouse used consists of Datamarts. The Datamart refers to a subset of the Data Warehouse that contains Data Warehouse data for a particular area. We speak for example of Datamart marketing, Datamart commercial, etc.

Furthermore, the database design of the proposed web scraping system is done in two stages. The first stage is to design the database using star modeling. Star modeling is a way of connecting dimensions and facts in a Data Warehouse. The principle is that dimensions are directly related to a fact (schematically, this is done like a star). The dimensions are the axes with which we want to analyze.

The facts, in addition to the dimensions, are what the analysis will focus on. These are tables that contain operational information (region, city, time, etc). Thus, the needs analysis led to the design of the logical data model shown in [Figure 4]. This schema appears like a star and connects the dimensions and facts.

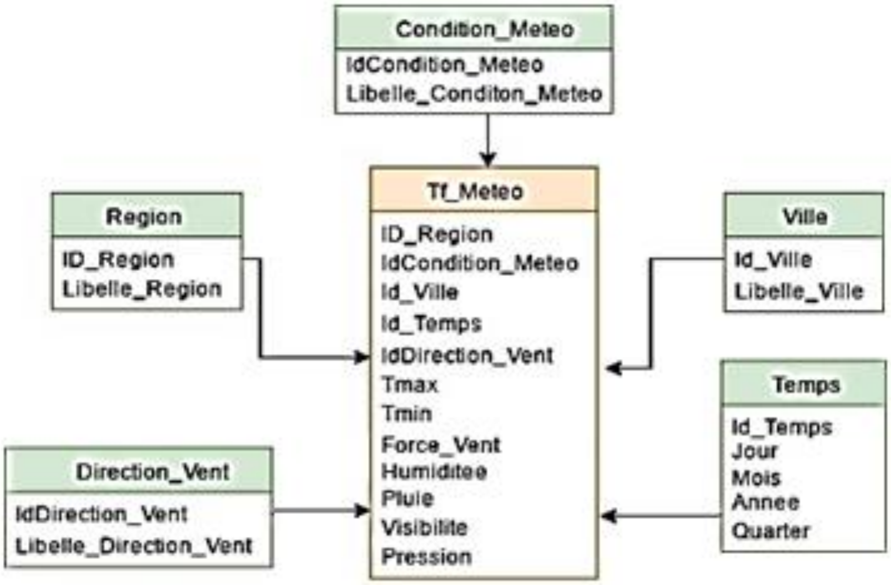


Figure 4. Web scraping framework database presented as logical data model

The database of the proposed Web scraping system integrates the following tables:

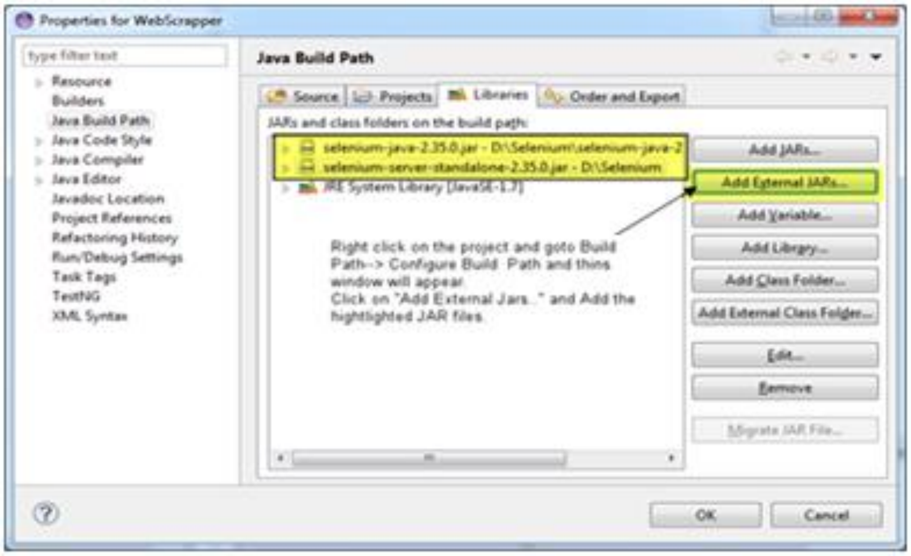
- Region: contains the different regions covered by the study;
- Direction_Vent: contains all possible wind directions;
- Condition_ Meteo: contains the nature of cloud cover (fairly overcast, some clouds);
- Temps: contains a history of each day's data;
- Ville: contains the names of the cities involved in the study.
- TF_ Meteo: it is the fact table that contains the measurements for the analysis.

After giving the logical data model, the second stage in the database design process consists of transforming the logical schema describing a Big data source into a NoSQL logical schema using an MDA (Model Driven Architecture) transformation process.

3.3. Implementation and experimentation

To implement the proposed Web scraping system, the choice of technologies is based on the reuse and the linking of many existing software solutions. Thus, we have used Java as a development language, Apache to setting up an HTTP server, and MySQL to implement the relational version of the database. On the other hand, the Talend tool is used as an Editor for software specialized in the integration of Big Data. Finally, the Qlik Sence tool is used as a platform to transform data into knowledge.

To configure the development environment, the libraries Selenium client & Webdriver language binding for Java are integrated into the used IDE (Eclipse). [Figure 5] illustrates the development environment configuration.



JARs and class folders on the build path:

selenium-java-2.35.0.jar - D:\Selenium\selenium-java-2
selenium-server-standalone-2.35.0.jar - D:\Selenium
JRE System Library [JavaSE-1.7]

Right click on the project and goto Build Path--> Configure Build Path and thins window will appear
Click on "Add External Jars.." and Add the hightlighted JAR files

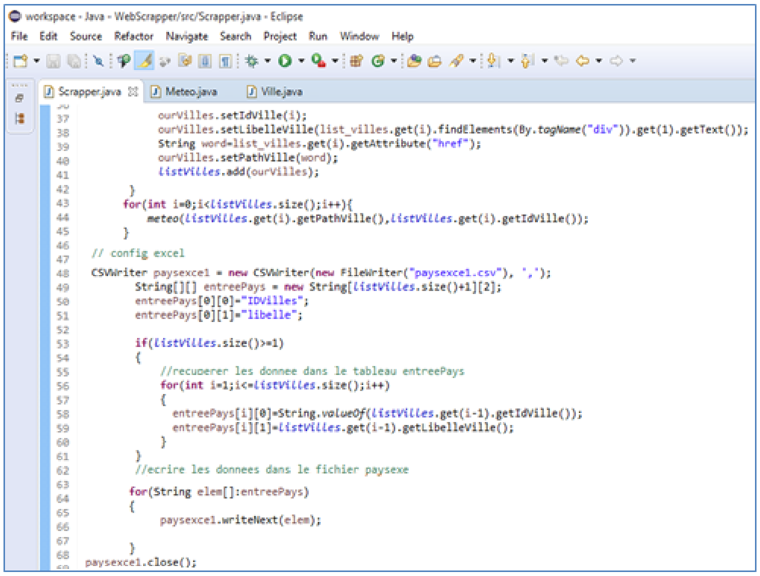
Figure 5. Selenium configuration

To experiment with the proposed framework, the meteorological data is taken from the following website: http://www.meteoma.net. This is a weather inventory site that presents weather data in Morocco for more than 260 towns and villages. To be able to scrap the data from this website, the source code of the website page must be downloaded to properly analyze where the data and tags are. This is facilitated by the inspector of a user web browser.

The analysis of the Web page structure allows us to develop the Java program to extract the necessary data using the following libraries:

- org.openqa.selenium.By
- org.openqa.selenium.WebDriver
- org.openqa.selenium.WebElement
- org.openqa.selenium.firefox.FirefoxDriver
- com.opencsv.CSVWriter

[Figure 6] illustrates a Java code fragment for extracting data on cities from the web page used (http://www.meteoma.net).



```java
                ourVilles.setIdVille(i);
                ourVilles.setLibelleVille(list_villes.get(i).findElements(By.tagName("div")).get(1).getText());
                String word=list_villes.get(i).getAttribute("href");
                ourVilles.setPathVille(word);
                listVilles.add(ourVilles);

            }
          for(int i=0;i<listVilles.size();i++){
              meteo(listVilles.get(i).getPathVille(),listVilles.get(i).getIdVille());
          }

     // config excel
     CSVWriter paysexcel = new CSVWriter(new FileWriter("paysexcel.csv"), ',');
            String[][] entreePays = new String[listVilles.size()+1][2];
            entreePays[0][0]="IDVilles";
            entreePays[0][1]="libelle";

            if(listVilles.size()>=1)
            {
                //recuperer les donnee dans le tableau entreePays
                for(int i=1;i<=listVilles.size();i++)
                {
                  entreePays[i][0]=String.valueOf(listVilles.get(i-1).getIdVille());
                  entreePays[i][1]=listVilles.get(i-1).getLibelleVille();
                }
            }
            //ecrire les donnees dans le fichier paysexe

          for(String elem[]:entreePays)
          {
                paysexcel.writeNext(elem);

          }
    paysexcel.close();
```

Figure 6. Java code for extracting data on cities

After the data extraction phase, the data is processed using the Talend Open Studio tool and sent to the system's Data Warehouse.

Once the data collection and extraction are done, these data are loaded to be viewed by the user in a simpler and clearer format. In this work, data generation and representation are done under a Dashboard using the Qlik Sense tool allowing to load of collected data into a Data Warehouse. The data loading leads to the generation of a data model illustrated in [Figure 7] below.

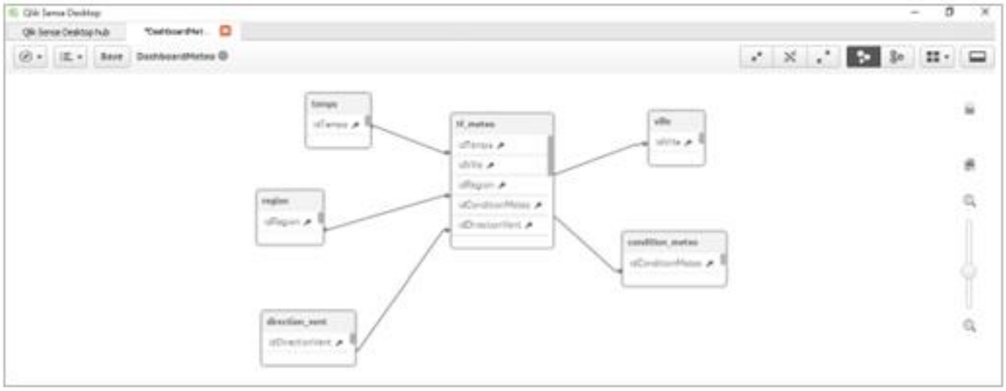


Figure 7. Data model generated after data loading

4. Results and discussion

In the implemented web scraping framework, the dashboard module generates the visualizations thanks to the Qlik Sense tool. In this interface, several types of graphs and indicators are used for the descriptive analysis of meteorological data. Colors are used in the graphs by dimension: i.e. more the color is darker, the more the value is great. The sheet Qlik Sense of [Figure 8] presents an example of descriptive meteorological data which includes five cities with maximum humidity, in several dimensions, adding the notion of filter, where the user can filter according to several criteria such as region, city, weather condition, wind direction, year, month, quarter, days, etc.

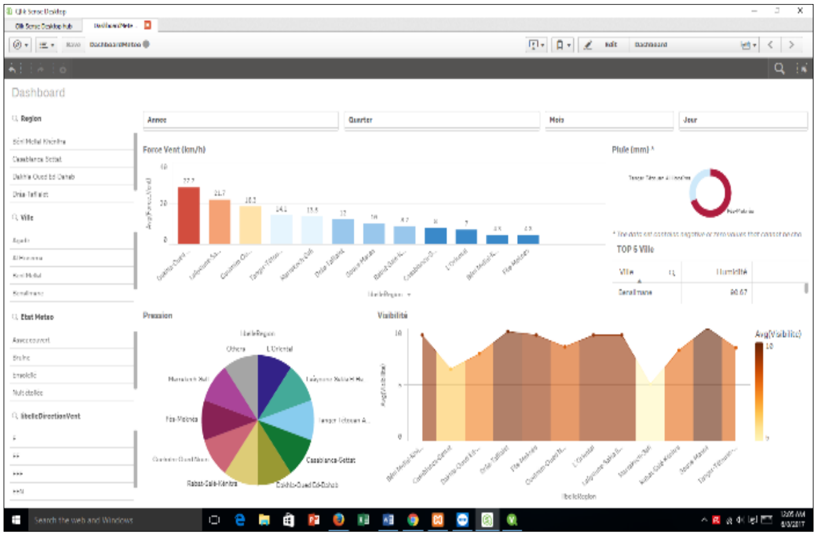


Figure 8. Sample Qlik Sense sheet to visualize data

Note that in this work, the automation aspect is managed by the operating system that will allow scheduling the execution of programs. Thus, the Java jar file is generated firstly for the extraction of data and Talend jobs. Then, the Java program is set to run every day at 00h:30min.

As a result, this interface makes it possible to produce reliable estimates and to make statistical inferences about the weather and the impacts that may occur over the season. For example, this data makes it possible to estimate the zones where the temperature is likely to be the highest or the lowest. They also help estimate the most likely amount of rain that will fall in a specific area.

Indeed, this descriptive analysis of meteorological data is essential for emergency planning. With this information, local authorities can better anticipate weather-related problems before they occur. For example, planners can make preparations to evacuate low-lying areas that are susceptible to flooding from heavy rainfall. It is also possible to make plans to modernize existing installations.

Consequently, data and indicators obtained through the dashboard allow to offer various analytical capabilities on huge amounts of data and help users to make optimal decisions related to the weather forecast. Furthermore, the proposed approach establishes a guideline for researchers and practitioners on how to analyze meteorological big data.

Based on the results of this study, and considering that weather can have a significant impact on many activities in people's daily lives, the proposed solution can bring many benefits in terms of prediction and decision support related to weather conditions. Today, for more accuracy in terms of predictions and decision support about weather conditions, people usually use several weather data analysis tools. The proposed solution is a contribution in this direction.

Indeed, weather forecasts play a very important role in early warning of weather impacts on various aspects of human life. During the last years, with emergent advancements in information technologies and data science, it is possible to take advantage of readily available numerical weather data issued from various sources to rapidly develop analytical and predictive models capable of predicting certain weather-related aspects.

In this context, many users only wish to base their decisions on a single forecast, but others, who have more specific needs, are interested in other possibilities. Based on this constatation, the present work has introduced a framework of descriptive analysis of meteorological big data based on web scraping techniques. The proposed framework, implemented using Java language and the Qlik Sence platform, allows the collection of meteorological big data and offers various data analytical capabilities and statistical models used for prediction and decision support purposes. Data files from the scraping process are transformed into a NoSQL database and stored in a data warehouse.

The results of the data collected by the web scraping techniques can then be presented as statistical models. Data statistics are created using the Qlik Sense platform. Data can be presented as needed by performing the data grouping process. For analysis and decision support purposes, data can be presented in the form of graphs and tables. Other forms of statistical data in this study are also presented in the form of a histogram. With the histogram presentation, it is easier for users to see the statistics of weather data of each city to observe. Additionally, the histogram can be presented in daily, monthly, and yearly forms as needed using the group function by the Date variable.

The study presented is a preliminary work at the stage of extraction and analysis of meteorological big-data issued from Web sources. Therefore, the process of descriptive analysis of meteorological data collected is limited to the presentation of statistical models for decision support regarding weather forecasts. However, this work establishes a guideline for researchers and practitioners on how to analyze meteorological big data and obtain accurate and timely weather forecasts. Moreover, the solution can be generalized to other domains.

Indeed, this work has detailed the analytical process and the presentation of statistical data for decision-support purposes related to weather forecasts, but it has not addressed the mapping of automatic prediction models. Thus, the data collected is still relatively weak. The data cannot be used to make automatic weather forecasts since the forecasting process itself ideally uses a lot of weather data to produce accurate forecasts. With the successful application of data-driven machine-learning techniques in various fields, it has been proven that the machine-learning method can effectively exploit weather features. This is why one of the research directions that we plan to pursue in the future is to continue the data collection process with web scraping techniques. Once enough data is collected, the next stage of research will be weather forecasting based on the machine learning approach.

A final aspect that concerns researchers in this field are the legality and fair use of the use of web scraping techniques which are often problematic. In this context, there are two aspects to consider in web scraping techniques, namely: copyright and unauthorized seizure.

5. Conclusion

Web scraping is a familiar technique that has gained importance because of the need to exploit data stored on web pages. Today, many researchers and professionals need the data to process it, analyze it and extract meaningful results or integrate it into new applications that provide added value and innovation. In this context, this paper has presented a Web scraping system that exploits Web scraping techniques for meteorological big data extraction and reuses it in complex processes.

Thus, this work has sought to develop a Web scraping system as a Java application allowing the analysis and visualization of data in a Dashboard and the filtering of such data to analyze it and use it for decision support purposes with weather forecasts. The paper has presented the design of the proposed system. To test its applicability, the proposed Web scraping system was implemented and tested in the meteorological context to extract and present meteorological data issued from http://www.meteoma.net/ web site that concern the north of Morocco. In this context, the different tools and technologies used in the development phase of the project are described.

The outputs generated by the proposed system are statistical and dynamic meteorological data models published in a dashboard. These data make it possible to predict certain meteorological variables (humidity, precipitation, temperature) and also to make decisions in terms of vigilance.

The possibilities of taking advantage of the proposed approach are numerous, including the descriptive analysis of meteorological data and the decision-making with weather forecasts. Beneficiaries can be people who only wish to base their decisions on a single forecast, but also other people who have more specific needs and who are interested in other possibilities. On the other hand, the proposed framework represents a brief guide for researchers who intend to develop similar systems either in the meteorological field or in other fields generating large amounts of data.

As part of the continuity of this work, it is planned to continue the data collection process with web scraping techniques and the use of data mining and machine learning approaches to enable the system to do the precise weather forecast.

Furthermore, this search work will continue with the experimentation of the Web Scraping system in another context, especially, the extraction and processing of data issued from e-learning environments, and this starting from the assumption that the analysis of learning data can identify the problems that students face and therefore improve the learning process.

References

[1] O. Castrillo-Fernández, “Web Scraping: Applications and Tools,” European Public Sector Information Platform Topic Report no. 2015 / 10, **(2015)**

[2] G. Lazyan, T. Baghdasaryan, and G. Aghajanyan, “The use of Big Data in Central Bank of Armenia,” Proceedings of the IFC-Bank Indonesia Satellite Seminar on “Big Data” at the ISI Regional Statistics Conference, Bali, Indonesia, March 21, **(2017)**

[3] A. Virgillito and F. Polidoro, “Big Data Techniques for Supporting Official Statistics: The Use of Web Scraping for Collecting Price Data." Web Services: Concepts, Methodologies, Tools, and Applications,” Edited by Information Resources Management Association, IGI Global, pp.728-744, (2019) DOI: https://doi.org/10.4018/978-1-5225-7501-6.ch040

[4] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol.19, no.2, pp. 171–209, (2014) DOI:https://doi.org/10.1007/s11036-013-0489-0

[5] P. Mikalef, I.O. Pappas, J. Krogstie and P. A. Pavlou, “Big data and business analytics: A research agenda for realizing business value,” Information & Management, vol.57, no.1, (2020) DOI: https://doi.org/10.1016/j.im.2019.103237

[6] D. Ushakov, E. Dudukalov, E. Mironenko, and K. Shatila, “Big data analytics in smart cities’ transportation infrastructure modernization,” Transportation Research Procedia, vol.63, pp.2385-2391, (2022) DOI: https://doi.org/10.1016/j.trpro.2022.06.274

[7] Y. Ye, J. Shi, D. Zhu, L. Su and J. Huang, Y Huang, “Management of medical and health big data based on integrated learning-based health care system: A review and comparative analysis,” Computer methods and programs in biomedicine, vol.209, no. C, (2021) DOI: 10.1016/j.cmpb.2021.106293

[8] M. A. Ashaari, K. S. Singh, G. A. Abbasi, A. Amran, and F. J. Liebana-Cabanillas, “Big data analytics capability for improved performance of higher education institutions in the Era of IR 4.0: A multi-analytical SEM & ANN perspective,” Technological Forecasting and Social Change, vol.173, no.1, (2021) DOI: https://doi.org/10.1016/j.techfore.2021.121119

[9] Q. Wang and Z. Mu, “Risk monitoring model of intelligent agriculture Internet of Things based on big data,” Sustainable Energy Technologies and Assessments, vol.53, part C, (2022) DOI: https://doi.org/10.1016/j.seta.2022.102654

[10] A. Aboutaoufik, “Big data in Morocco's transport and logistics sector,” International Journal of Innovations in Engineering and Technology, vol. 19, no. 2, pp.27-35 **(2021)**

[11] M. Lamrabet, T. Benkaraache, “Big data et systèmes décisionnels au Maroc : État des lieux,” Economie Digitale et PME en Afrique, Casablanca, Maroc, Post-Print hal-03463064, HAL **(2019)**

[12] P. Kaur, “Sentiment analysis using web scraping for live news data with machine learning algorithms,” Materials Today: Proceedings, vol. 65, part 8, pp.3333-3341, (2022) DOI: https://doi.org/10.1016/j.matpr.2022.05.409

[13] M. J. Lee, J. Kang, K. Hreha, and M. Pappadis, “A Novel Web Scraping Approach to Identify Stroke Outcome Measures: A Feasibility Study,” Archives of Physical Medicine and Rehabilitation, vol. 103, no.3, (2022) DOI: https://doi.org/10.1016/j.apmr.2022.01.082

[14] E. Vargiu and M. Urru, “Exploiting web scraping in a collaborative filtering based approach to web advertising,” Artificial Intelligence Research, vol.2, no.1, pp. 44-54, (2013) DOI: https://doi.org/10.5430/air.v2n1p44

[15] E. Ensari and B.Kobas, “Web scraping and mapping urban data to support urban design decisions,” AZ ITU Journal of the Faculty of Architecture, vol.15, no.1, pp.5-21 (2018) DOI: 10.5505/itujfa.2018.40360

[16] Y. Adhinugroho, A. Putra, M. M. Luqman, G. Y. Ermawan, T. Takdir, S. Mariyah, and S. Pramana, “Development of online travel Web scraping for tourism statistics in Indonesia,” Information Research: An International Electronic Journal, vol. 25, no.4, (2020) DOI: 10.47989/irpaper885

[17] J. D. Ford, S. E. Tilleard, L. Berrang-Ford, M. Araos, R. Biesbroek, A. C. Lesnikowski, G. K. MacDonald, A. Hsu, C. Chen and L. Bizikova, “Big data has big potential for applications to climate change adaptation,” Proceedings of the National Academy of Sciences of the United States of America, vol.13, no.39 (2016) DOI: 10.1073/pnas.1614023113

[18] R.S. Chaulagain, S. Pandey, S.R. Basnet and S Shakya, “Cloud Based Web Scraping for Big Data Applications,” Proceedings of the IEEE International Conference on Smart Cloud (SmartCloud), New York, USA, November 3-5, **(2017)**

[19] E.G. Ularu, F.C. Puican, A. Apostu, and M Velicanu, “Perspectives on Big Data and Big Data Analytics,” Database Systems Journal, vol. 3, no.4, pp.3-14, **(2012)**

[20] R. Kune, P.K. Konugurthi, A. Agarwal, R.R Chillarige and R. Buyya, “The anatomy of big data computing,” Software: Practice And Experience, vol.46, pp.79–105, (2016) DOI:10.1002/spe.2374

[21] I. K. Nti, J. A. Quarcoo, J. Aning, and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," Big Data Mining and Analytics, vol.5, no.2, pp.81-97, (2022) DOI: 10.26599/BDMA.2021.9020028

[22] M.R. Trifu and M.L. Ivan, “Big Data: present and future,” Database Systems Journal, vol.5, no.1, pp.32-41, **(2014)**

[23] T. White, “Hadoop: The Definitive Guide, 4th Edition,” O'Reilly Media and Inc. Publishers, **(2015)**

[24] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker and I. Stoica, “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,” Proceedings of

the 9th USENIX conference on Networked Systems Design and Implementation, San Jose CA, USA, April 25 - 27, **(2012)**

[25] Apache flink. http://flink.apache.org/.

[26] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, “Distributed GraphLab: a framework for machine learning and data mining in the cloud,” Proceedings of the VLDB Endowment, vol.5, no.8, pp.716–727, (2012) DOI: 10.14778/2212351.2212354

[27] J. Rogers, R. Simakov, E. Soroush, P. Velikhov, M. Balazinska, D. DeWitt, B. Heath, D. Maier, S. Madden, J. Patel, M. Stonebraker, S. Zdonik, A. Smirnov, K. Knizhnik, Paul G. Brown, “Overview of SciDB: Large scale array storage, processing, and analysis," Proceedings of the International Conference on Management of Data, Indianapolis, Indiana, USA, June 6-11, **(2010)**

[28] M. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches, and Application," International Journal of Advances in Soft Computing & Its Applications, vol.13, no.3, pp.144-168, (2021) DOI:10.15849/IJASCA.211128.11

[29] K. Vasani, “Content Evocation Using Web Scraping and Semantic Illustration,” IOSR Journal of Computer Engineering, vol.16, no.3, ver.IX, pp.54-60, (2014) DOI:10.9790/0661-16395460

[30] A. Simitsis and P. Vassiliadis, "Extraction, Transformation, and Loading," In Liu, L., Özsu, M.T. (eds) Encyclopedia of Database Systems, Springer, New York, NY, **(2018)**

[31] R. Vording, “Harvesting unstructured data in heterogenous business environments; exploring modern web scraping technologies,” Proceedings of the 34th Twente Student Conference on IT, Enschede, Netherlands, January 29, **(2021)**

[32] E. Persson, “Evaluating tools and techniques for web scraping,” M.S. thesis, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, Stockholm, Sweden, **(2019)**

[33] R. N. Landers, R. C. Brusso, K. J. Cavanaugh and A. B. Collmus, “A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research,” Psychological Methods, vol.21, no.4, pp.475-492, (2016) DOI:10.1037/met0000081

[34] I. Onyenwe, E. Onyedinma, C. Nwafor, and O. Agbata, “Developing Products Update-Alert System for e-Commerce Websites Users Using HTML Data and Web Scraping Technique,” International Journal on Natural Language Computing, vol.10, no.5, pp.1-8, (2021) DOI: 10.5121/ijnlc.2021.10501

[35] H. Kasereka, “Importance of web scraping in e-commerce and e-marketing”, SSRN Electronic Journal, pp.1-10, (2021) DOI:10.6084/m9.figshare.13611395.v1

[36] M. Herrmann and L. Hoyden, “Applied Web scraping in Market Research,” 1st International Conference on Advanced Research Methods and Analytics, Valencia, Spain, July 6–7, **(2016)**

Authors

Abderrahim El Mhouti
Received the Ph.D. degree in Computer Science in 2015 from Abdelmalek Essaadi University, Tetouan, Morocco. Mr. EL MHOUTI is a Professor of Computer Science at the Faculty of Sciences at the same University. His research fields include Big Data and Learning analytics, cloud computing, educational technologies, and machine learning. He has published several articles in his areas of research.

Mohamed Fahim
Received the Ph.D. degree in Computer Science from Moulay Ismail University, Meknes, Morocco. Mr. Fahim is a Professor of Computer Science at the Faculty of Sciences and Technologies of Al-Hoceima belonging to Abdelmalek Essaadi University. His research fields include technology-enhanced learning, educational technologies, machine learning, and deep learning. He has published several articles in his areas of research.

Adil Soufi
Received the Ph.D. degree in Computer Science from Abdelmalek Essaadi University, Tetouan, Morocco. Prof. Soufi is a Professor of Computer Science at the Faculty of Sciences and Technologies of Al-Hoceima at the same university. His research fields include machine learning, e-learning, modeling, and fitting for epidemic models. He has published several articles in his areas of research.

Imane El Alama
Received an engineering degree in mathematics and computer science. Ms. Imane is passionate about IT and new technologies, particularly in the field of Business Intelligence and Big Data. She is responsible for the design and development of several projects around the implementation of computer systems, Java development, and decision support systems.