Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 2, (Special Issue) April-2018, ISSN 2230-7540

Web Scraping-Data Extraction Using Java Application and Visual Basics Macros

Sameer Padghan[1]* Satish Chigle[2], Rahul Handoo[3]

[1,2,3] JSPM's Imperial College of Engineering and Research, Wagholi, Pune

Abstract: ***There is a lot of data on the internet. The data is increasing exponentially and even the data on the internet is in unstructured format. The semi-structured and unstructured data make it difficult to extract relevant data from the Web. Even if the data is extracted properly the efficiency is quite low which make the data irrelevant. To answer this problem there comes in picture the Web Scraping. Web scraping can be defined as the extraction of data from the website, in general term. The project with the help of web scraping will make the data extraction from the website easily. This approach will allow scraping the data from various website, which will reduce the human effort, save the time, also increase the data relevancy efficiency. This will help the user to extract data from the website and save the data for his purpose and use as the user wants it. The scraped data can be used to create database or for analysis purpose as well as for various other purpose.***

Keywords: Web Scraping, Data extraction, visual basics macros, structured data

----------X----------

1. INTRODUCTION

[1] Web scraping usually mean extraction of data from the website. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding and artificial intelligence and human-computer interactions. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that are able to convert entire web sites into structured information, with limitations.

[2] Web Harvey says "Web Scraping (also termed Screen Scraping, Web Data Extraction and Web Harvesting etc.) is a technique widely used to extract large amount of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format. Data displayed by most websites can only be viewed using a web browser. Examples are data listings at yellow pages' directories, real estate sites, social networks, industrial inventory, online shopping sites, contact databases etc. Most websites do not offer the functionality to save a copy of the data which they display to your computer. The only option then is to manually copy and paste the data displayed by the website in your browser to a local file in your computer - a very tedious job which can take many hours or sometimes days to complete." The concept of Web Scraping is not new to us, it is getting more famous these days because of the new Startups, as they don't have to do much hard work to get the data, they mostly prefer to use the data scraped from other similar websites and then they modify it as per their need. Due to this, the bigger existing companies are facing much loss as their data being anonymously gathered and then reproduced by some other companies.

This was generally done manually where user needs to click the website links and then further copies the required data from the websites and it is pasted in a spreadsheet that could be on local computer or on the Google sheet but the process web scraping automates all the work. There were already existing systems that used to extract data those are Scrapy, Web data extractor, Outwit hub etc. The problem in the previously used application was that it was not user friendly and cannot be used by a general user further he needs to learn about the core application and the coding part too.

A web scraping software will automatically load and extract data from multiple pages of websites based on your requirement. It is either custom built for a specific website or is one which can be configured to work with any website. With the click of a button you can easily save the data available in the website to a file in your computer. The problem with most generic web scraping software is that they are very difficult to setup and use. There is a steep learning curve involved.

[2] Web Harvy was designed to solve this problem. With a very intuitive, point and click interface, using

Web Harvy you can start extracting data within minutes from any website. We all use Web Browser to extract the needed information from the Web Sites, if you think this is the only way to access information from internet then you are missing out a huge range of available possibilities. Web Scraper is one of those possibilities, in which we are accessing the information on the internet using some programs and pre written libraries.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process, so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time. Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. Web scraping a web page involves fetching it and extracting from it .WSAPI is the platform that enables an organization to extend their existing web based system, as well designed set of services for creating new channels, developer integration or partner integration. It helps to offer clean and structured data from existing websites, so that the data can be effortlessly consumed by disparate systems. The data that is being exposed through these APIs can be monitored, transformed and controlled easily. The inherent design helps developers to incorporate website changes without affecting the extraction logic by moving them to configurations.

2. LITERATURE SURVEY

Initially the extraction of data was done manually by humans, but now with the help of modern programing language it is possible to do automate the process. But even after the availability of such thing it is not quite use due to sophisticated interface which is not user friendly. Also few times the data after extraction is not relevant the quality is not up to the mark. There are issues even with some website, they do not allow web scraping and they block such kind of things. They do not allow web scraping to any extent. Generally during the scraping, data is extracted only form a particular website, but we aim to do it from the various website which can be given as a input in a file or a keyword in the application. This will ease the process of manual extraction and will reduce redundant work of the user. There were already existing systems that used to extract data those are Scrapy, Web data extractor, Outwit hub etc. The problem in the previously used application was that it was not user friendly and cannot be used by a general user further he needs to learn about the core application and the coding part too.

In Web scraping data is extracted from the websites that can be in any format. Web scraping a web page involves fetching it and extracting from it.Fetching is the downloading of a page source code (which a browser does when you view the page).Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).It can be used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mash up and, web data integration.

A) Title- R-Extractor: A Method for Data Extraction from Template-Based Entity-Pages

In this context, the domain-centric data extraction (DCDE) methods arose through replacing user intervention with content redundancy. The DCDE methods extract the attribute values of entities of the web that are restricted to a specific application domain.

B) Title-Agent Mat: Framework for data scraping and semantization

Web data extraction is the process of extracting user required information from websites. From the word web data extraction, we mean the extraction of data that is present in the web documents in HTML format. In this paper, we have studied about different techniques for data extraction used by different authors that takes the user required data from a set of web pages. A comparative analysis of web data extraction techniques is given.

c) Title-Web data extraction techniques

Web data extraction is the process of extracting user required information from websites. The web document contains data which is not in structured format. From the word web data extraction, we mean the extraction of data that is present in the web documents in HTML format.

D) Title-DEPTA:

An efficient technique for web data extraction and alignment

This paper studies the issue of extracting these data records from online web database. The main motto of this paper is to recognize the data region which contains the data records, divide these data records, mine the data value from them and keep these extracted record in a structured format.

E) Title-Data pre-processing algorithm for Web Structure Mining:

This paper based on the first two stages Data collection and preprocessing. Data

Collection is to collect the data required for analysis. Data preprocessing is considered as an important stage of Web Structure mining because of data.

3. WEB SCRAPING TECHNIQUES

Human copy and paste: This is the traditional way used for web scraping in which the user copies the required data per his need and it is the only resolution in the case where automation fails

Text pattern matching: In this techniques UNIX grep command or regular expression matching facilities are being used which is one of the powerful approach to extract data.

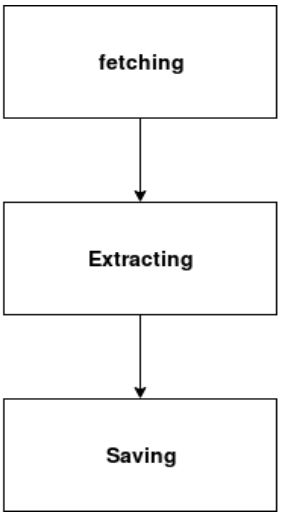
HTTP programming: HTTP requests can be remotely send using socket programming in which static and dynamic pages can be retrieved.

HTML parsing: Every website has a structured source like a database which has large collection of pages that are being dynamically generated .In data mining, a program that detects such templates in a particular information source, extracts its content and translates it into a relational form, is called a wrapper.

Computer vision web-page analysis: Efforts are being made for machine learning and for enhancing computer vision to identify and extract information from webpages that can be interpreted by the user.

4. PROPOSED SYSTEM

The system has three parts namely fetching, extracting, saving to local database.



Fetching: The process start when the user gives the website, links of website or the keyword in the application. In the case, when the sole website link is given it fetches it the web page directly. After this the extraction of the data takes place from the fetched web page. While in the case of list of website same strategy took place, only in the case of the keyword the application get the link from the search engine and then start the usual procedure of extraction of the data.

Extraction: The second part is the extraction of the data, in this part the Visual basic module does the job of getting the relevant data form the system.

Saving: The last but not the least, the saving of the data to native database

In these cases the data extracted from the fetched web page is being saved to the spreadsheet or in a pdf file or can be saved in a database as per the user requirement.

Diagram. The source code on the web is being copied in the spreadsheet by the web scraping application and then the relevant data is extracted from the source code with the help of the visual basic macros. The visual basic macros extract the data, then the data is being saved in the spreadsheet or in any PDF format. In java code there are built libraries that allow web scraping from HTTP, HTTPS, and XML etc. Macros are being written in visual basics which are further being used in MS-excel

.

In this architecture, the web scraping is dichotomized half on online process and half on the offline process. Here online process means that you need a internet connection with proper bandwidth and further offline tool is used to extract information from the data that is received after exploring the source code .*Firstly*, JAVA application is used to extract source code from the website which is pasted in a spreadsheet. *Secondly*, visual basic macros are being used to extract relevant information from the source code and further this data is shown to the user in a structured form. This data can be used for contact making and lead generation. Further databases can be formed on this basis. This data can be used for demographic analysis, resource analysis etc. The working of web scraper is quite simple, it starts with a list of URL to visit which are called seeds. The web scraper will try to scrape data but yet there is some website that does not allow scraping which make it impossible to scrape them. The web scraper then fetches the source code of all the web pages given in the list only of website which allow scraping. From this source code the main process of extracting will start afterward. This source code of this entire website is saved in the spreadsheet. Then the visual basic macros run on this source code. The visual basic macros then start to extract the relevant data from the source code. The extracted data is saved either in the spreadsheet or it is saved in the PDF file format or it can be saved in the database. The visited website link source code is available to us from which we will extract the relevant data. Looking for the whole website will be worthless so only the relevant information is converting as per the user need following steps is given as follow:

- Finding the content on search engine by entering the appropriate term which will make getting website quite easy for extraction.
- Avoiding the recurring website, i.e. the particular website will be taken only one time.
- Extracted data is needed to be stored in a well formatted manner so it can be reusable which is possible by using CSV format.
- Saving the data in CSV format help us to edit the data quite easily than any other format.
- After having the data in CSV format which is quite among everyone now day, the user will be able to convert it into any other format as per his need.

In this implementation we are using JAVA as coding language, the reason behind choosing JAVA is that ,it has vast community support and enough libraries such as **Jsoup**.[5]Jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

The visual basic macros are used to get the relevant data from the source code of the website. After getting the data from the source code the visual basic support the function for saving data in the CSV or various other format. The data is generally saved in the PDF document. The extracted data can be used for various purposes as per the user requirement. In this web scraping is done.

5. IS SCRAPING WEBSITE ALLOWED?

In this case the question always remains unanswered. Actually it depends upon the user from where he is extracting the data. If he is extracting data from the server and he is not an authorized person to access the server then it is illegal. But in case he is extracting data that is available on the website, then it is legal because the person who is providing data on the page of the website wants people to extract that data for lead generation and contact making.

6. FUTURE SCOPE

Those days are not away when everyone will use start using web scrape. There are various service such as price comparison, big data analysis, lead generation that can be used by masses. Since the data is increasing gigantically on the internet the web scraping will be very prominent in the cutting edge technology. With the help of modern programming language like Python, R, Ruby, Scala customized web scraping development will be very easy in future. The web scraping will provide a solution for all those who want data for their big data analysis.

7. CONCLUSION

The next generation can be manipulated using web scraping. It is a modern generation tool with which anyone can know about the activities, events, friends and location etc. of any random person. “It is always the small pieces that make big pictures”. From the discussed clauses the conclusion is that the used of scraping will increase drastically and it can sometimes intrude the system to fetch the information. But by using proper anti-web scraping techniques this extraction can be avoided. This tool should be considered as a boon that should be wisely used for the development of human race.

8. REFERENCES

DEPTA: An Efficient Technique for Web Data Extraction and Alignment 2010 WASE International Conference on Information Engineering

HTML Web Content Extraction Using Paragraph Tags Howard J. Carey, III, Milos Manic Department of Computer Science Virginia Commonwealth University Richmond, VA USA

http://webscraper.io/

https://en.wikipedia.org/wiki/Web_scraping

https://jsoup.org/

https://www.quora.com/topic/Web-Scraping

https://www.webharvy.com/articles/what-is-web-scraping.html

Parallel Approach and Platform for Large-scale Web Data Extraction,2013 International Conference on Advanced Cloud and Big Data

Ryan Mitchell – Web Scraping Using Python, First Edition, Orilley,June 2015

Corresponding Author

Sameer Padghan*

JSPM's Imperial College of Engineering and Research, Wagholi, Pune

E-Mail – sameerpadghan@gmail.com