International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
363 – 367

An Overview On Web Scraping Techniques And Tools

Anand V. Saurkar
Assistant Professor
[1]Department of Computer Science & Engineering
[1]Datta Meghe Institute of Engineering, Technology & Research, Swangi(M), Wardha, Maharashtra, India
saurkaranand@gmail.com

Kedar G. Pathare
Assistant Professor
[1]Department of Computer Science & Engineering
Datta Meghe Institute of Engineering, Technology & Research, Swangi(M), Wardha, Maharashtra, India

Shweta A. Gode
Assistant Professor
Department of Computer Technology
Yeshwantrao Chauhan College of Engineering, Nagpur, Maharashtra, India
shweta_amt80@rediffmail.com

***Abstract*—** From the evolution of WWW, the scenario of internet user and data exchange is fastly changes. As common people join the internet and start to use it, lots of new techniques are promoted to boost up the network. At the same time, to enhance computers and network facility new technologies were introduces which results into automatically decreasing in cost of hardware and website's related costs. Due to all these changes, large number of users are joined and use the internet facilities. Daily use of internet cose in to a tremendous data is available on internet. Business, academician, researchers all are share their advertisements, information on internet so that they can be connected to people fastly and easily. As a result of exchange, share and store data on internet, a new problem is arise that how to handle such data overload and how the user will get or access the best information in least efforts. To solve this issues, researcher spotout new technique called Web Scraping. Web scraping is very imperative technique which is used to generate structured data on the basis of available unstructured data on the web. Scaping generated structured data then stored in central database and analyze in spreadsheets. Traditional copy-and-paste, Text grapping and regular expression matching, HTTP programming, HTML parsing, DOM parsing, Webscraping software, Vertical aggregation platforms, Semantic annotation recognizing and Computer vision web-page analyzers are some of the common techniques used for data scraping. Previously most user uses the common copy-pest technique for gathering and analyzing data on the internet, but it is a tedious technique where lot of data copied by the user and store on computer files. As compared to this technique web scraping software is easiest scraping technique. Now a days, there are lots of software are available in the market for web scraping. Our paper is focused on the overview on the information extraction technique i.e. web scraping, different techniques of web scraping and some of the recent tools used for a web scraping.

Keywords- Web mining, information extraction, web scraping

*****

I. INTRODUCTION

In the field of marketing, scientific or academic research data plays an important role. Researcher, market analyzer or academicians gather data from different websites for their better improvement. Copping of data on the website to user local storage in forbidden by most of the website authority. So that the user want to manually coping the data from website to local computer file storage. But such a task is very exhausting and time consuming. Due to such limitation web scraping techniques are introduces. By using web scraping techniques user can extract information available on multiple website into a single database or spreadsheets. So data can be easily visualize and analyse for further use. Web scraping technique is a sub-discipline of web mining technology. The main objective of this paper is to review on the different web scraping techniques and software which can be used to extract required information from web sites.

Web Mining stays at the crossroads of Information Retrieval, Information Extraction and Data Mining. Both Information Retrieval and Information Extraction have important roles to trace and mine valuable information out of unstructured data, before it is suitable to be processed by data mining applications. Exploring better these techniques is extremely necessary to cope with the amount of available data in the Information Overload Era. Also, with the Web being increasingly oriented towards the importance of semantics and integration of information, these areas of study become very important to address the new future trends of the Web.

Figure 1: web mining [3]

Information Retrieval (IR) is a field of study concerned with the retrieval of documents from a collection of other documents (relevant and non-relevant), usually based on key word searches. With the Internet expansion, Information Retrieval got a very special focus, as search engines became a dominant way to find information on the Web. Nowadays, because of

 ISSN: 2454-4248
363 – 367

their importance, search engines are the most representative name in Information Retrieval.

Information Extraction (IE) is a sub-part of artificial intelligence. IE mainly focus on to extract valuable data out of unstructured data. An extraction information system is usually aimed on entities or objects identification (people, places, companies, etc) and extraction rules. Videos, images, audio and text are some of the common example of unstructured data.. The first information extraction systems were mainly focused on text, and still nowadays this is the most explored type of data by the research community and commercial frameworks. The goal of IE is to identify useful parts out of raw data (unstructured data) and extract them to finally build more valuable information through semantic classification. The result may be suitable for other information processing tasks, such as IR and Data Mining. Characteristically there is a difference between the goals of IR and IE, but in the actual world they should be seen as close complementary activities to improve their precision and accuracy[3].

It is important to highlight that the term extraction define that the information which user want is explicitly available . The need for this explanation is to differentiate between statistical data mining techniques which deduce information out of available data. However the information retrieval and information extraction can build knowledge from a given data set, it does not mean that they can be used alternatively. Actually, information extraction is a necessary pre-processing step to structure data before a statistical data mining algorithm can build knowledge (at this point still hidden) from it. To extract a useful and important data from a retrieved information different techniques are used. Among those techniques, in this paper we are focus on the web scraping techniques and discuss on the some of the tools which are available for web scraping [3].

II. OVERVIEW OF WEB SCRAPING

Web Scraping is important technique used for extracting unstructured data from the websites and transforming that data into structured. Web Scraping is also identified as web data extraction, web data scraping, web harvesting or screen scraping. Web scraping is a form of data mining. The basic and important aim of the web scraping process is to mine information from a different and unstructured websites and transform it into an comprehensible structure like spreadsheets, database or a comma-separated values (CSV) file. Data like item pricing, stock pricing, different reports, market pricing and product details, can be gathered through web scraping. Extracting targeted information from websites contributions to take effective decisions in business process.

Figure 2. Basic Architecture of Web Scraping

As we know, scraping is a technique used to crop information from web pages based on script routines. Web pages are documents written in Hypertext Markup Language (HTML), and more recently XHTML which is based on XML. Web documents are represented by a tree structured called the Document Object Model, or simply the DOM tree and the goal of HTML is to specify the format of text displayed by Web browsers as shown in figure 3.

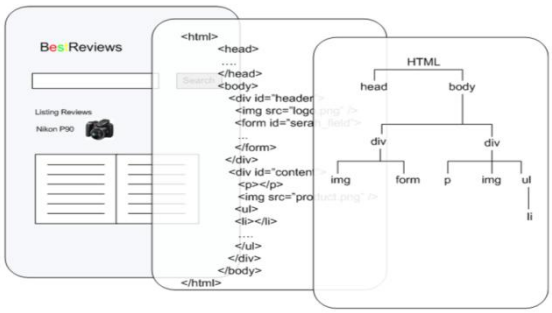


Figure 3: Three different outlook of web document - The document on web, the HTML code and the Document Object Model [3]

From the operation viewpoint, a web scraping look like manual copy and paste task. The difference here is that this job is done in a organized and automatic way, by a virtual computer agent. When an agent is following each link of a web page, it is actually performing the same operation that a human being would normally do when interacting with a web site. This agent can follow links (by issuing HTTP GET requests) and submit forms (through HTTP POST), browsing through many different web pages. While a computer would perform manual tasks at the speed of a computer instruction, a human would have to think, grab the mouse, point to the link and finally click on it. Now the benifit seems clear when a user has to click on several links before getting to the actual desired page. However, not surprisingly, the benefit of issuing requests at a script speed, also brings a problem. If one uses web scraping without a policy for limiting requests, the requested server may find that someone is trying a Denial-of-Service attack, due to the great amount of requests triggered in a short period of time [3].

Next step is, the parser follows user-specified paths inside the document to retrieve the desired information based on the data retrived in previous step. These paths are specified by CSS selectors or XPATHs. They use both relative and absolute paths (based on the DOM tree) to point the parser to a specific element inside a web document. Normally web

 ISSN: 2454-4248
363 – 367

scraping operations uses regular expressions to narrow or trim the located information, in order to retrieve data with an user-specified granular size. This process is illustrated in figure 4[3].A web scraping agent gathering information from web pages - The dotted circle represents a web scraping agent traversing a DOM tree. The red lines are XPATHs to a desired element within the document. The agent reaches the hyperlink in web page 1 and proceeds to web page 2 until it finds the information enclosed by element p (paragraph).

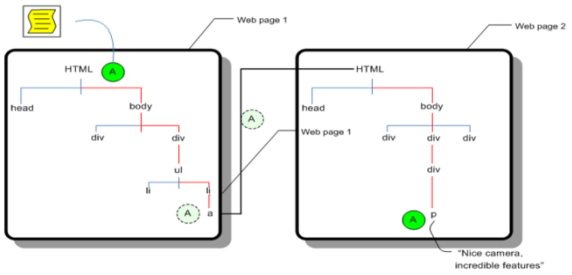


Figure 4: Last step of web scraping- Gathering information by scraping agent on the basis of DOM tree and XPATH [3].

III. WEB SCRAPER USES

Web Scrapers are also being used by Online Marketers to pull data privately from the competitor's websites such as high targeted keywords, valuable links, emails & traffic sources. Some of the area where web scraper techniques are mostly uses are:

- Change detection on website
- Product Price comparison
- Weather broadcasting and data monitoring
- Research analytics
- Analyze data in graphics
- Web Indexing & rank checking
- Advertisement analysis
- Market Analysis

IV. WEB SCRAPING TECHNIQUES

A. Classical copy and paste: The human's manual examination and copy and-paste method is the best and the workable webscraping technology. But this is an tending to implement or cause errors, and tiresome technique when user need to analyse and store lots of datasets[1].

B. Hypertext Transfer Protocol (HTTP) Programming: By using this technique user can be extract information from static and dynamic web pages. Data can be retrieved by posting HTTP requests to the remote web server using socket programming .

C. Hyper Text Markup Language (HTML) Parsing: Semi-structured data query languages, like XQuery and the Hyper Text Query Language (HTQL), can be used to parse HTML pages and to retrieve and transform page content [1].

D. Document Object Model (DOM)Parsing: By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages.

E. Web Scraping Software: Now a days many software tools are available, that can be used to customize web-scraping solutions. This software may attempt to automatically recognize the data structure of a page or provide a recording interface that removes the necessity to manually write web-scraping code, or some scripting functions that can be used to extract and transform content, and database interfaces that can store the scraped data in local databases.

F. Computer vision web-page analysers: There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might [1].

V. WEB SCRAPING TOOLS

Web Scraping Software are the computerized program that are used to make the manual copy paste work automatically .it also collect large amount of information from websites like directory sites, real estate sites, classified websites and job boards and stored on local server. Suppose you want to scrape real estate property details of India then you need to appoint few guys to copy and paste details from websites to excel by visiting each property agent pages. This way it will take days and even months to get your property data ready to use. So web scraping can automate the manual work programmatically by visiting each page and extract data from pages and parsing the html pages. There are number of Web Scraping Software that available in market that can help you to scrape data from any website you want. Following are the list of some scraping tools [1].

Web scraping software work as, to bot or web crawler access the web data directly using the Hypertext Transfer Protocol, or through a web browser and extract the precise data from that web page. This extracted data is then store into a central local database or spreadsheet for later use or analysis.

Some of the tools are :

5.1 Mozenda:

It is a Business Intelligence Software. It is allows user to extract data from documents as the same way user can extract data from web page. Also user can be combine data from multiplesourses into a single data set. Mozenda currently

 ISSN: 2454-4248
363 – 367

support documents scraping for several popular formats like World, Excel, PDF etc. key features of Mozenda are: industry Data Feeds, high-volume weekly data feeds, project building and project maintenance. Mozenda will automatically detect names and associated values and build robust data sets with minimal configuration. Mozenda's Data Mining Software is packed full of useful applications especially for sales people. User can do things such as "lead generation, forecasting, acquiring information for establishing budgets, competitor pricing analysis. This software is a great companion for marketing plan & sales plan creating. Using Refine Capture tex tool, Mozenda is smart enough to filter the text you want stays clean or get the specific text or split them into pieces.

5.2 Visual Web Ripper :

Visual Web Ripper is one of the most advance web scraping software, created by Sequentum group in 2006. It offers more functionality which will allows user to scrape data from any websites like Business Directories, Simple Web Pages, Classified Sites, Forums and e-commerce site like eBay, amazon, magento sites. After finishing data scraping task data can be exported to structured CSV, Excel, or XML format.

***5.3** Web Content Extractor:*

Another Simple and user oriented tool for data scraping is Web Content Extractor (WCE) which is developed by Newprosoft. It has good wizard that guide user to setup scraper. User can scrape data from website with few clicks and WCE is self intelligent for putting data into different formats like Excel, text, HTML formats, Microsoft Access database, Structured Query Language(SQL) Script File, MySQL Script File, Extensible Markup Language (XML) file, HTTP submit form and Open Database Connectivity (ODBC) Data source.

***5.4** Import.io*

Import.io is a web based tool for extracting data from website without writing code. If user want a fast result then he/she will try for this tool so that he can convert website in short time. For extracting data, user enter URL and application automatically extract data which user want to need, if user does not interested in the automatic extraction, the point and click interface help to select data fields on website. As the data extraction is over, the extracted dataset is store on Import.io cloud server and ferther downloaded in CSV, Excel, JSON format.

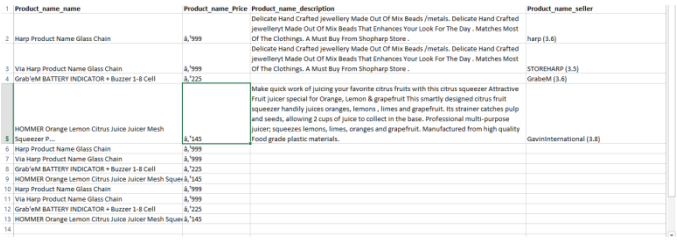
***5.5** Scrapy:*

An open source and collaborative framework for extracting the data you need from websites. Scrappy is designed to scrape web content from sites that are composed of many pages of similar semantic structure. The system is implemented as a Firefox browser extension, and works in three main stages to scrape web data. First, a user navigates to a page that he would like to scrape and generates a template for the content that he would like from that page. Next, the user selects a set of links that point to pages matching the content template defined by the user. Finally, the user selects an output data format and Scrappy crawls the links specified by the user and scrapes content corresponding to the user's template[9].Scrapy written in Python and runs on Linux, Windows, and Mac.

VI. OVERVIEW ON STUDY OF SOME SCRAPING TOOL.

We have a number of tools available for scraping a web data, some of them are briefly describe in above section. For a experimental analysis of how to perform a scraping using web scraping tools, we study two tools. Firstly we study 'scraper', a web scraping extension available in chrome web store. It is easily downloaded and apare in chrome tool bar. Scraper is a very simple but limited data mining extension for faciliting online extraction of data for researcher in the format of spresdsheets. Scrapy gives a result in the form of spreadsheet for a website which we select for scraping. Data is in limited format I.e. we can not get proper data in spreadsheet.

After this we study for the free available scraping tool 'ParseHub'. It a simply point and click software base on a standard API. User have full controle on extraction of data from targeted website. It work like a hierarchical base selection of data. At the starting of scraping, user simply select the field which he/she want to extract, then ParseHub automatically guess similar data element from a website. As user select a related information which he want to extract al similar element are extracted. For selecting a other data elements from a targeted website a 'relative' search option is available which is subset information about previously select element. Likewise user extract all information from website. At the time of extraction of element from a website, ParseHub provide a URL also. This URL is optional field. After successful web scraping data sets are saved in a CVS format as shown following.



	Product_name_name	Product_name_Price	Product_name_description	Product_name_seller
2	Harp Product Name Glass Chain	₹,'999	Delicate Hand Crafted jewellery Made Out Of Mix Beads /metals. Delicate Hand Crafted jewellery! Made Out Of Mix Beads That Enhances Your Look For The Day . Matches Most Of The Clothings. A Must Buy From Shopharp Store .	harp (3.6)
3	Via Harp Product Name Glass Chain	₹,'999	Delicate Hand Crafted jewellery Made Out Of Mix Beads /metals. Delicate Hand Crafted jewellery! Made Out Of Mix Beads That Enhances Your Look For The Day . Matches Most Of The Clothings. A Must Buy From Shopharp Store .	STOREHARP (3.5)
4	Grab'eM BATTERY INDICATOR + Buzzer 1-8 Cell	₹,'225		GrabeM (3.6)
5	HOMMER Orange Lemon Citrus Juice Juicer Mesh Squeezer P...	₹,'145	Make quick work of juicing your favorite citrus fruits with this citrus squeezer Attractive Fruit juicer special for Orange, Lemon & grapefruit This smartly designed citrus fruit squeezer handily juices oranges, lemons , limes and grapefruit. Its strainer catches pulp and seeds, allowing 2 cups of juice to collect in the base. Professional multi-purpose juicer; squeezes lemons, limes, oranges and grapefruit. Manufactured from high quality Food grade plastic materials.	GavinInternational (3.8)
6	Harp Product Name Glass Chain	₹,'999		
7	Via Harp Product Name Glass Chain	₹,'999		
8	Grab'eM BATTERY INDICATOR + Buzzer 1-8 Cell	₹,'225		
9	HOMMER Orange Lemon Citrus Juice Juicer Mesh Squee	₹,'145		
10	Harp Product Name Glass Chain	₹,'999		
11	Via Harp Product Name Glass Chain	₹,'999		
12	Grab'eM BATTERY INDICATOR + Buzzer 1-8 Cell	₹,'225		
13	HOMMER Orange Lemon Citrus Juice Juicer Mesh Squee	₹,'145		
14				

Figure 5: Data Set after scraping online scraping tool

VII. CONCLUSION

To get automatic information from website, web scraping is the most effective and efficient technique. Among all other techniques mention in this paper which are used to extract and

 ISSN: 2454-4248
363 – 367

store data, web scraping is more reliable, fast and automic data retrieval system. By using web scraping terminology user can easily extract unstructured data on single or multiple websites into a structured data automatically. The main aim of this technique is to get information from web and aggregate into a new dataset. In this paper we are discuss the basic of web mining. We also focus on the techniques used for web scraping. The last section of the paper promote the overview of different software tools available in market for proper web scraping.

REFERENCES

[1] S.C.M. de S Sirisuriya,2015, A Comparative Study on Web Scraping .Proceedings of 8th International Research Conference, KDU.

[2] List of Web Harvester, Data Scraper,Web Scraping Software and Tools, n.d.WebData Scraping. URL http://webdata-scraping.com/webscraping-software/

[3] Felipe Jordão Almeida Prado Mattosinho, Master Thesis, Mining Product Opinions and Reviews on the Web, TU Dresden.

[4] Eloisa Vargiu, Mirko Urru1, 2013, Exploiting web scraping in a collaborative filtering- based approach to web advertising, Artificial Intelligence Research, 2013, Vol. 2, No. 1, http://dx.doi.org/10.5430/air.v2n1p44.

[5] Schrenk, M. Webbots, spiders, and screen scrapers: a guide to developing Internet agents with PHP/CURL. No Starch Press, 2007.

[6] OsmarCastrillo-Fernández, 2015, Web Scraping: Applications and Tools, European Public Sector Information Platform, Topic Report No. 2015/10.

[7] Deepak Kumar Mahto, Lisha Singh, A Dive into Web Scraper World, 2016 InternationalConference on Computing for SustainableGlobal Development (INDIACom), 978-9-3805-4421-2/16/$31.00 c , 2016 IEEE.

[8] MiloslavBeiio, Jakub Misek, Filip Zavoral, AgentMat: Framework for Data Scraping and Semantization, 9781-4244-2865-6/09/$25.00 ©2009 IEEE.

[9] Amir Ghazvinian, Sean Holbert, Nikil Viswanathan, Scrappy: Simple Web Scraping, Department of Biomedical Informatics, Department of Computer Science, Stanford University.