Web scraping approach in an API world

Sneha Y. Bhosale[1], Pournima U. Kunnure[2], Amruta R. Kamble[3], Snehal U. Dhamnekar[4], Komal B. Sawant[5], Prof. A. S. Yadav (Guide)[6]

[1,2,3,4,5] UG students at D.Y. Patil College of engineering and technology, Kolhapur
[6]D.Y. Patil College of engineering and technology, Kolhapur

Abstract- *Daily use of internet cause 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 spot out 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. Scarping 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, Web scraping 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. The main objective of this paper is to offers a review on web scraping using simple HTML DOM parser.*

Keywords- Web scraping, data mining, web mining, and information extraction.

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 wants 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. 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 from it. To extract a useful and important data from retrieved information different techniques are used.

In proposed system, we consider all reviews from different websites like Amazon, Flipkart, Snapdeal. This reviews are useful for analyze the product is good or bad and then the final rating of each product is given by our system based on review analysis Review extraction is done by using this web scraping technique. Web scraping is best suited as alternative to API. This information is useful for customer to choose the product. Based on above analysis the system is also recommend to seller. The review analysis is in particular time span. The graph representation is used to display the result for sellers. So, to increase their profit our system is useful to choose the product which is being sold.

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.

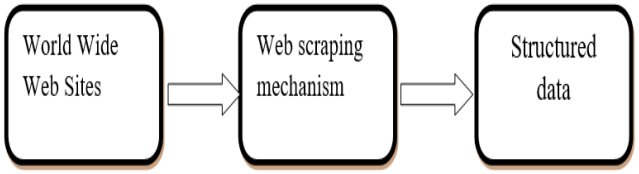

Fig 1. Basic Architecture Web Scraping

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. 2

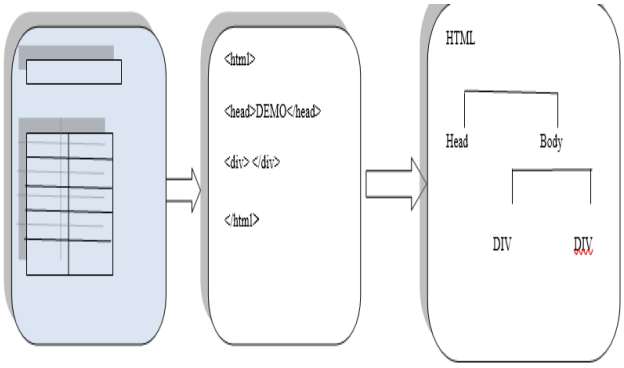

Fig. 2 Three different outlook of web document - The document on web, the HTML code and the Document Object Model

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.

Next step is, the parser follows user-specified paths inside the document to retrieve the desired information based on the data retrieved 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.

III. RELATED WORK

First we fire URLs, according to url scraper script designed. Scraper mainly containing simple DOM parser in HTML which fetches to particular div data by using div class name in fired URL. Fetched data will be stored in database at structured format for further use.

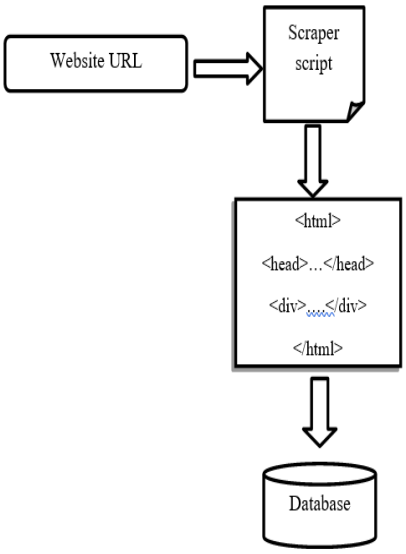

Fig. Data extraction using web scraping overview.

IV. CONCLUSIONS

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 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. Now we can conclude that use of Scraper in the coming world will be increased significantly. As Scraper opens up another world of retrieving information without the use of API, and mostly it is anonymously accessed.

REFERENCES

[1] Jian Jin ,PingJi , RuiGu "Identifying comparative customer requirements from product online reviews for competitor analysis" Elsevier 2016.

[2] Vedita velingker, Malony alphonso" Recommender System based on product reviews" , IJSTE june 2016.

[3] 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.