Web Data Extraction From Websites For Pricing Analysis

Charu Sharma Student
Department of Master of Technology,
Himachal Pradesh University, Shimla, India.

1.Abstract

Web Extraction or Scraping is the process of extracting data from the internet. Today, internet usage has been greatly increased for collecting information and so the amount of data on it. This results in increasing competition among various organizations which are running their businesses on the internet. Web Scraping can be done either by using ad-hoc methods which requires human effort or by web scrapers which are fully automated software that automatically convert the web data into structured form for future analysis of the data. The usage of Web Scraping is in many fields and is beneficial for weather data monitoring, website change detection, Research, Web Data Integration, Contact scraping and Online Price Comparison. This paper focuses on one of the objective of Web Scraping i.e. Pricing Analysis. Pricing Analysis is the process of analyzing the variation in prices of a product among various organizations. The paper describes use of automated software to collect the data from websites for price analysis. The software described is Octoparse which will collect the data from three different websites i.e. L.G., Samsung and Sony. This paper will compare the prices of different products of Television from each website i.e. UHD, OLED, QLED and LED.

2.Introduction

With the use of web as a source of wide range of information the availability of data on the web has increased manifolds so the complexity of the data as well. Today, there are many organizations who are running their businesses on the internet which has increased the use and analyzation of the data on the web to improve their businesses.

Web Extraction in simple terms is to collect data from the internet for future analysis and review of it[2]. There are many softwares available which scrape the data from the web and convert it to a structured form for further analysis. Thus, Web Extraction is the primary task in Web Mining. The data on the web is unstructured and for analysis the data required should be in structured form. Hence, various softwares or ad-hoc methods are available for converting the unstructured data into a form which can easily be used for future analysis of the content[3].

Web Mining as the name suggests is the process of applying various mining techniques on the web data. Web Mining is a broad field and thus is divided to three broad categories depending upon the data mined. The three categories are:- Web Content Mining, Web Structure Mining and Web Usage Mining[9].

In the modern era where use of internet and online trading is increased use of web data extraction has a wide range of applications. The areas where web extraction is used are:- Marketing, Trading, Research and Analysis[2][4].

3.Pricing Analysis

Pricing Analysis is one of the important area which uses web extraction. Pricing Analysis is done mostly by market or trading analysts to understand the existing behavior of the market for future predictions. Online trading in the market is done by Pricing Analysis of the web data.

To understand the pricing statistics in the era where the data available is highly complexed and with the rise in competition of online trading various automated softwares are available. These softwares scrape the prices of the content available on the web and convert them to a structure which can easily be further used to analyse the price variation of the data for future predictions or decisions[1].

4.Tool of Web Extraction

There are many softwares available which extract the unstructured data on the web and store the data on a spreadsheet for future use. The softwares are made from languages such as Java Script, PHP, Python, etc. Softwares convert the extracted data into an API which can be easily used for further processing[3]. Use of such softwares has greatly reduced the need of manual work for collection of the data as now the task of collecting data has been automated with the use of softwares.

A software of web extraction performs five different functions to extract the data from the web. Firstly, the software navigates the web pages of the website from where the data is to be extracted. Secondly, the softwares used for extraction has programs known as wrappers which provides a GUI interface to the user to identify the data which is required from the web page. Thirdly, once the data is selected by the user wrappers will navigate the action repeatedly in a loop for the next similar web pages from the site in order to capture the desired data. Fourthly, softwares transforms the data into the desired format which requires filtering, mapping, refining, and integrating of data from one or more sources. Fifthly, the software used will transport the data structured to some database or a spreadsheet for future analysis[4].

There are number of softwares used for web data extraction and these softwares are available both free and commercially. This paper focuses on the tool – **Octoparse.**

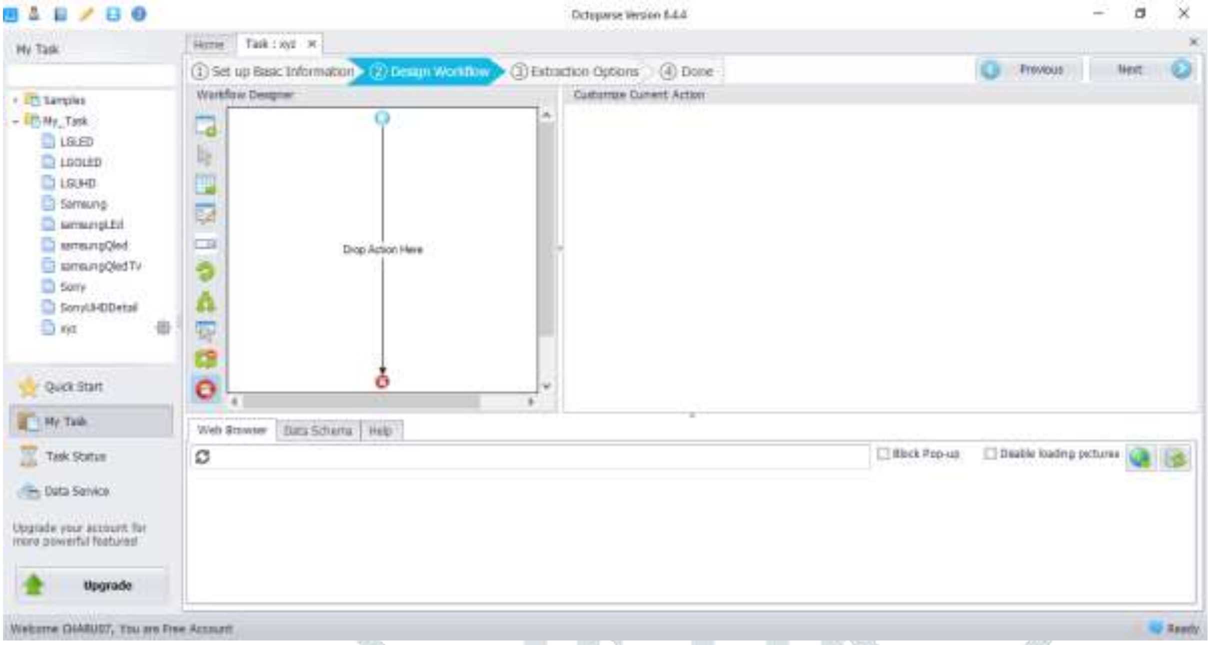
4.1 Octoparse

Octoparse is free software of web data extraction. Ocotparse provides a most user friendly GUI for the user to easily extract the data from all websites.

There are many unique features of the tool. The list of such features is given below:-

1. Point-and-Click Interface
2. Deals with all websites
3. Cloud Extraction
4. Automatic IP Rotation
5. Scheduled Extraction
6. API, CSV, Excel, Database
7. Free[10]

The general interface of the tool is shown below:-



Octoparse Interface – **Figure 1**

Procedure of Extraction from the tool – Octoparse

In order to correctly extract the data from any website, Octoparse has a defined approach to scrape. This approach has following steps:

1. Go to the web page – This is the first step where the user will enter the website link in the address bar of the software and click GO on the software. It will direct the user to the corresponding website.
2. Click an Item – This is the second step where the user will click an item on the page loaded which the use wants to extract. After clicking the software will provide the user with various actions that the user wants to perform further in the form of easily understandable manner.
3. End the workflow – This is the last step where the user will end the workflow of extraction once the user has selected the desired data needed from the website.

Once the data is selected and workflow is ended, user can extract the data using local extraction interface. Here, all the items selected will be displayed in a structured format i.e. tabular which can be easily exported to CSV, Database and Excel for future analysis of the content.

5.Scraping Websites Using the Tool

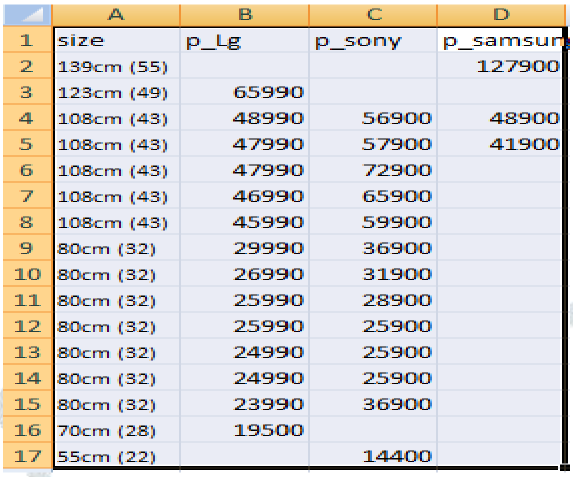
As each website has a wide categories of data such as text, image, hyperlinks, graphics etc., to collect this wide range efficiently a scraping software is used[1]. The tool used is Octoparse for data extraction from the websites.

This paper focuses on extracting the prices from various Television brands websites and of various different technologies from each website.

This paper uses a Web Extraction Tool – **Octoparse** to collect a **Price Data** of different categories of televisions i.e. **LED, UHD, OLED and QLED** from three official websites of television companies i.e. **Samsung, Sony and LG.** After the data is fetched using the tool and is transformed into a structured form and stored in a database, the data is then used to compare the prices of the three companies for each category of television.

6.Experiment Analysis

In this paper, using tool – Octoparse price data is collected using parameter of Screen_Size of each typed of television i.e LED, UHD, OLED and QLED.

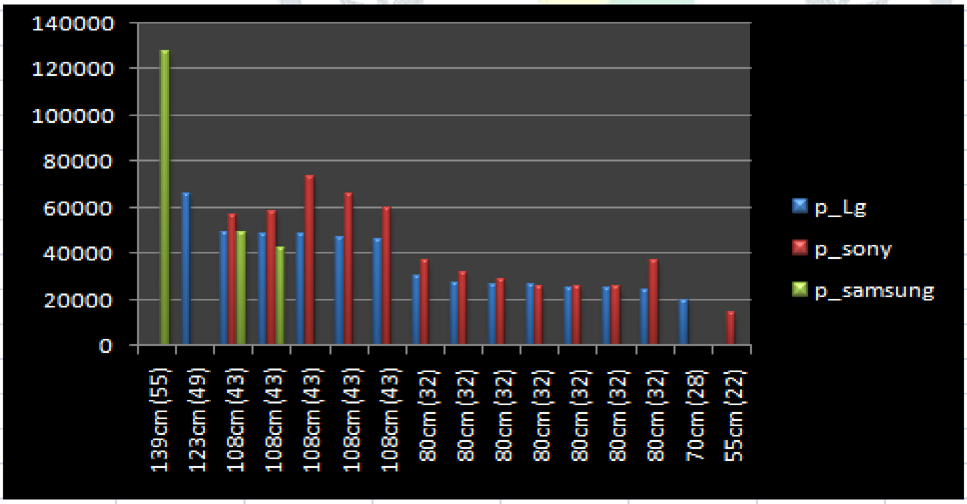


	A	B	C	D
1	size	p_Lg	p_sony	p_samsun
2	139cm (55)			127900
3	123cm (49)	65990		
4	108cm (43)	48990	56900	48900
5	108cm (43)	47990	57900	41900
6	108cm (43)	47990	72900	
7	108cm (43)	46990	65900	
8	108cm (43)	45990	59900	
9	80cm (32)	29990	36900	
10	80cm (32)	26990	31900	
11	80cm (32)	25990	28900	
12	80cm (32)	25990	25900	
13	80cm (32)	24990	25900	
14	80cm (32)	24990	25900	
15	80cm (32)	23990	36900	
16	70cm (28)	19500		
17	55cm (22)		14400	

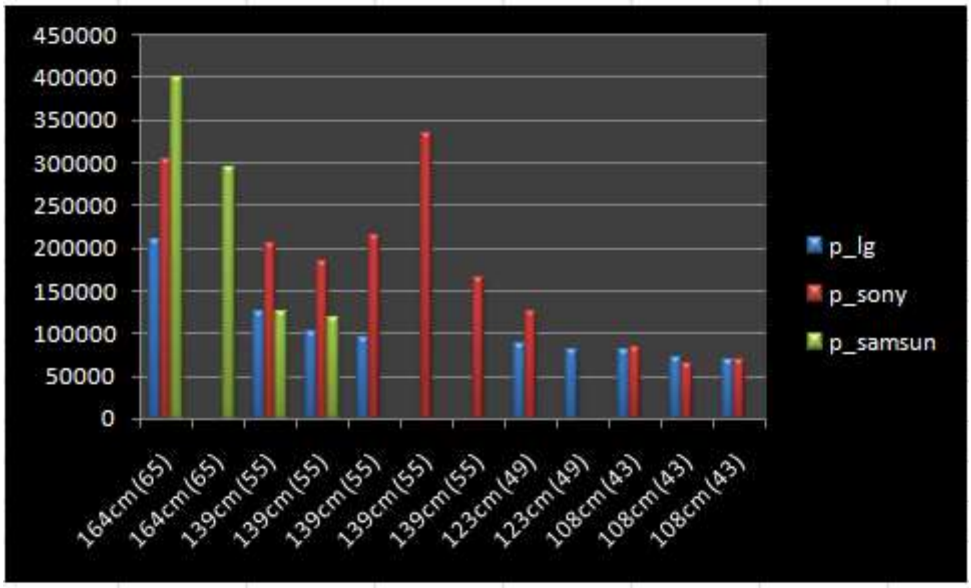
LED Data Set – **Figure 2**

Once the data set is collected and is stored for future analysis in a database, it can be used easily by either a user or a market analyzer to understand the price variation of its competitor. This paper analyzes price variation using Excel.

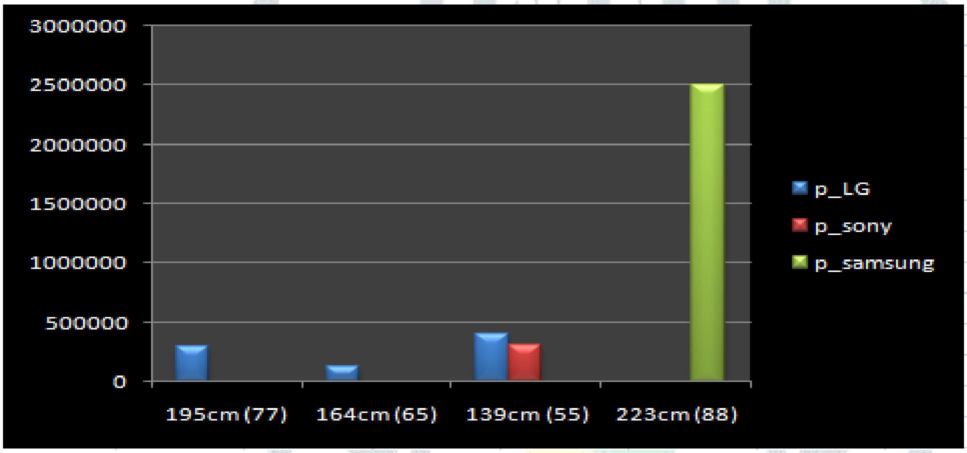
6.1 Variation of Prices



LED Price – **Figure 3**



UHD Price – **Figure 4**



OLED and QLED Price – **Figure 5**

7.Summary and Future Scope

Web Extraction is a process of scraping data from any website. Web Extraction forms the basis of every Web Mining Category and is mostly used recently due to increase in complexity of the data available on the web. This paper focuses on analysis of price variation of different television categories of different electronic companies. From the experiment analysis it is concluded that Samsung prices is high in all three categories of television in comparison to LG and Sony company as the screen size is more that the rest two.

The future scope of the work will be that using the analysis the firm can easily set the price of any new product. As in this world of competition, every firm should know what are the new launches and the prices of products of the another company which is in competition. This will help the firm to keep its position in the market . Also the work is beneficial for the user as well to understand the price variation before going for a purchase of a television of the electronic company.

References

1.Ingolf Boettcher - Automatic data collection on the Internet (Web Scraping) - VERSION 18 - May 2015.

2. Shikha Mahajan, Nikhit Kumar - A Web Scraping Approach in Node.js - International Journal of Science, Engineering and Technology Research (IJSETR) - Volume 4, Issue 4 - April 2015.

3. Renita Crystal Pereira, Vanitha T - Web Scraping of Social Networks - International Journal of Innovative Research in Computer and Communication Engineering - Vol. 3, Special Issue 7- October 2015.

4. Yolande Neil - Web Scraping the Easy Way - Digital Commons@Georgia University - 2016 .

5. Emilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd - Web Data Extraction, Applications and Techniques: A Survey – June 5,2014.

6. Jussi Myllymaki - Effective Web Data Extraction with Standard XML Technologies – 10 May,2001.

7. Eloisa Vargiu1, Mirko Urru - Exploiting web scraping in a collaborative filtering- based approach to web advertising - Artificial Intelligence Research - Vol. 2, No. 1 – 2013.

8. Nagesh Kumar Jha1, Aakash Jethva2, Nidhi Parmar3 , Professor Abhay Patil - A Review Paper on Deep Web Data Extraction using WordNet - International Research Journal of Engineering and Technology (IRJET) - Volume: 03 Issue: 03 - Mar-2016.

9. Shipra Saini, Hari Mohan Pandey –"Review on Web Content Mining Techniques" - International Journal of Computer Applications (0975 – 8887)- May 2015 -Volume 118.

10. https://www.octoparse.com/Product.