Exploiting Filtering approach with Web Scrapping for Smart Online Shopping

Penny Wise: A wise Tool for Online Shopping

Shakra Mehak
Department of Computing and Information Technology
Sialkot, Pakistan
University of Sialkot
shakramehak@gmail.com

Rabia Zafar
Department of Computing and Information Technology
Sialkot, Pakistan
University of Sialkot
rabiazafar600@gmail.com

Sharaz Aslam
Department of Computing and Information Technology
Sialkot, Pakistan
University of Gujrat, Sialkot Sub Campus
sherazaslam63@gmail.com

Sohail Masood Bhatti
Department of Computer Science
Sialkot, Pakistan
Government Women College and University
Sohail.bhatti@gcwus.edu.pk

***Abstract*— With the advancement in technology and popularity of e-commerce, the number of online shopping websites have been increased rapidly in the cyber world. This made people's life easy because it is easy to shop through internet. But this also bring effort for people as they spend a lot of time and efforts to search best product deals and offers on e-commerce websites. They have to filter and compare data by themselves. It takes a lot of time and still there are chances of ambiguous results. This paper is based on web crawling and scraping methods applied for identifying best deals from five e-commerce websites. The framework is designed using HTML (Hypertext markup language) and CSS (Cascading style sheet) as front-end and PHP: Hypertext preprocessor language as back-end support. The scrapping scripts are written using python libraries and web crawling works on HTML labels. The novelty in this framework is that we are not storing scrapped data on local database. Instead the results are dynamically fetched and showed every time the user input the query. It will help to improvise the storage and processing ability. Furthermore, the data retrieval process accuracy is 93% with minimum computation and less time.**

Keywords—Scraping, Crawling, HTTP, Web Driver, DOM

I. INTRODUCTION

The web world is extremely rich in terms of informative data and contents accessible in various formats like numbers, text, images, audio and video etc. on web pages which prompts irregularity in retrieval of information because of its insignificance for which the user searching for [1]. The massive amount of data can be blessing and offensive as well. The growing demand of data to urge individuals to grow new techniques and technologies with the goal that the access of data can be speedy and simple. The process of information retrieval is based on storing data on database, reverting information according to user query.

Nevertheless, 53.7% of Internet used by people looking for data about merchandise or administrations, 47.4% data is searched for educational purpose, 39% contents are searched for health and clinical data , 27.9% job seeking actions, 23.9% data are searched for governmental and law administrative organization [2]. However the action of web based shopping has turned into a smooth movement for web clients. The fulfillment of user getting to web based business locales are turning into a need with the end goal to expand online shopping deals and best priced products [3]. Online shopping is turning into core need of business now days because there is no need of planned shopping schedule and geographical constraints in online shopping [4]. Through e-commerce web sites user can see product details and buy tirelessly without visiting physical market. In spite of the fact, the low cost products with good qualities and time and energy saving are reasons that more individuals prefer online shopping. With the fame of Internet and online business, the number of shopping sites has quickly expanded on the Internet, and this empowers individuals to shop effectively through the Internet. Buyers invest tons of energy in looking feasible products, since they have to channel and think about list item's information without anyone else's input. Because of burdensome search, one needs to have a single platform that contains best deals available for a product on different sites.

Web scraping is an automated technique or software of extracting interesting data from websites. This method generally centers around to the change of unstructured or massive data (HTML/XML documents) on the web into organized information according to user query[5][6]. It is an incredible data mining technique for working with massive data on the web.

Web Scraping (likewise named Screen Scraping, Web Data Extraction, and Web Harvesting.) is a method employed to harvest a lot of information from sites and deployed to your local system[7][8]. The procedure of web scraping can be divided into two stages successively:

1) Obtaining web sources

2) extraction of relevant data from acquired sources[2].

Conventionally, this kind of programs simulated user exploration of web by executing hypertext transfer protocol

978-1-5386-9509-8/19/$31.00 ©2019 IEEE

(HHTP), or installing a completely fledged internet browser, for example, Google chrome[6]. Web harvesting is firmly identified with web indexing, which records data on the web using a bot or web crawler and is an inclusive method implemented in most web search tools[9] [10]. Additionally it stimulates user browsing using application software.

The idea of Web Scraping is not unfamiliar to us, as it is getting more renowned nowadays due to the new online business and Startups, as they don't need to do much diligent work to get the information. Preferably, they used scraped data from other similar sources and the change it according to their need. The applications of web scraping are generally observing weather data, research data gathering, seeking health information, and finding interesting patterns for business or web data integration[11].

II. METHODS

The system is implemented to build a website that search for product categories through HTML DOM-based architecture by using web scraping and web crawling strategies.

We used five e-commerce websites as information domain, Daraz.pk, GOTO, Telemart, Yeywo, CheezMall.

All these sites are analyzed on the basis of product categories. The web contents will be scrapped whenever user input query

A. Working Principle

The web application is implemented by following steps given below:

1. Import the Python libraries
2. Fetching the URL using request and selenium libraries and save it into temporary variable
3. Parse the HTML in temp variable and convert it into BeautifulSoup format
4. Scrap Product label, price, Specification and image
5. Compare the product price
6. View scraped data according to price

The 1st two steps come under the umbrella of web crawling that is done by using python libraries and 3rd and 4th steps are known as web scraping.

B. Web Scrapping/ Crawling Implementation using python

In this implementation we used python as coding language because python provides fast and powerful libraries and it offers community support for web scrapping and crawling.

We used request, Beautiful Soup 3 and Web Selenium Driver python libraries for scraping for different phases;

1. Python Requests

For opening and passing HTTP URLs python, request library is imported. Requests Library is a straightforward and simple to utilize HHTP library written in python. It makes communication with web services consistent and helps in connection pooling. This permits to continue parameters and treats over all requests that produced using the session occurrence. Python asks for encode the parameter automatically and interprets the response in Unicode[12]. On the other hand, multiple file sharing could be handled. Python request supports the entire Restful API, i.e. All your CRUD (CREATE, RETRIEVE, UPDATE & DELETE) methods.

2. Beautiful Soup 3

It is a library that fetches data of the web page either in HTML or XML format. In combination with parser it generates a parse tree based on DOM(Document Object Model) approach and then different filter functions can be applied to find particular tag, string, attributes or combination of all [13]. Let discuss the process firstly the document to be parsed is given as argument to beautiful soup function then the respective document is converted to UNICODE and elements in it become UNICODE characters. These characters are given as input to parser by default HTML parser runs however you can also specify the parser you want to use. When we talk about this library, it considers document and its elements as objects e.g. tags, navigable String, beautiful soup and comment [14].

3. Selenium

Selenium is best of its kind framework for testing, it provides full support to many browsers like Google, Firefox etc. It provides many testing operations for testing web applications. Selenium is a web driver that supports web pages that are dynamic in nature i.e. they have ability to support a page whose elements may change without reloading of page itself [15]. Although it has been used to compose web tests for web based application but it can be used for pages that have JavaScript on them. It is available in Python Web driver Selenium package[16].

C. Website Module

I-User Interface

We designed a simple and user friendly interface. By using this interface user can query by adding desired product name. The results of the query are product specification, image, Product price and name according to information source.

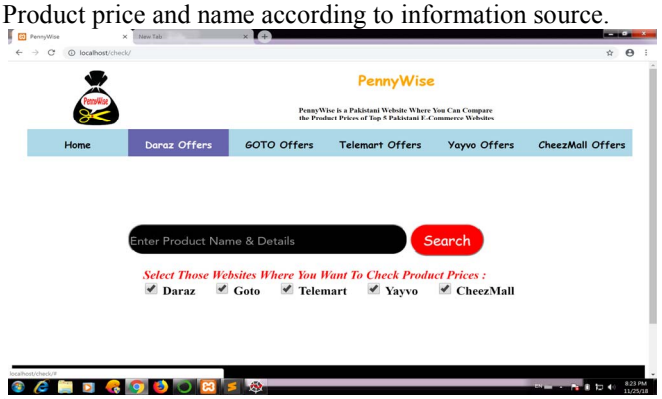

Fig. 1. User Interface

II-Business Logic

The application layer is responsible for retrieving selected attributes from website. It is based on special scripts that have been written in python and uses Beautiful Soup library to parse data. The web application is responsible for interacting with user and presenting them required results.

The working flow of web application and data scraping is shown in fig.1 and fig.2. The process starts by user generated query this query is embedded with URL to find at

least 6 lowest price instances of the particular product from each web domain and obtained the HTML labels. After finding instances there data and metadata including price, label, image and product specification are fetched for price comparison by using web scraping. The HTML labels that have been obtained from fetched page will be consequently parsed using HTML DOM (Document Object Model).

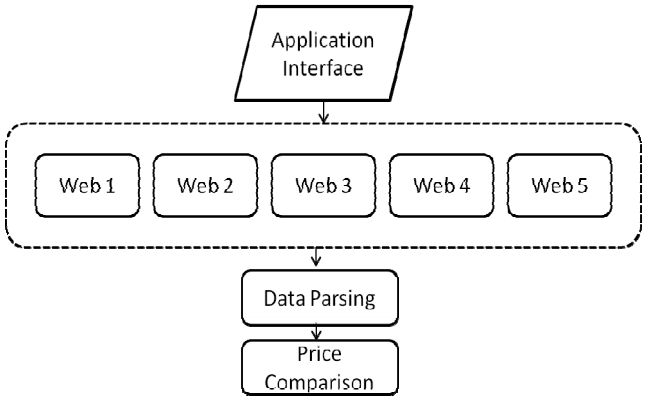

Fig. 2. System Overview

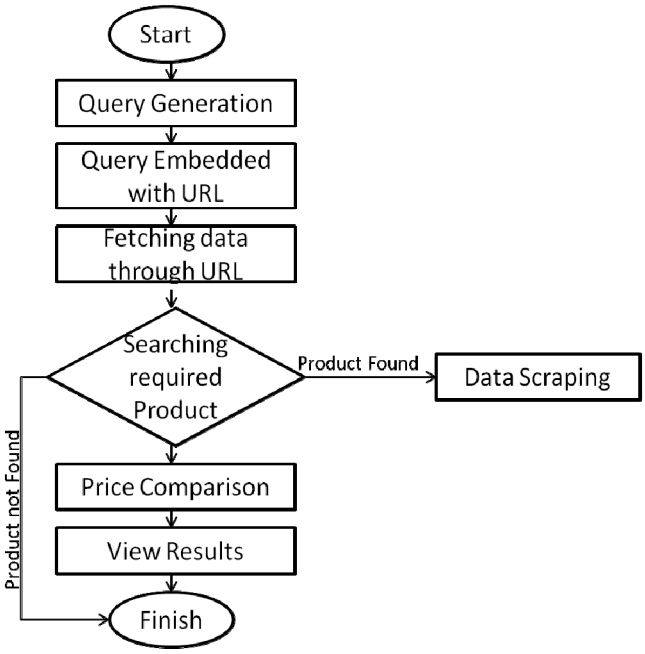

Fig. 3. Data Scraping Flow

DOM parsing utilized by including internet browser with the goal that the program can take dynamic substance produced from web content on the user side. Data that has been effectively parsed will be select with the labeling in each target site and compared set variable accordingly. Each instance taken from a web domain is compared with every other instance of other domains and finally 5 instances with lowest price of all are displayed as result of searched product on our web application.

III. Results and Discussion

This study implements a PHP based web application that provides facility of accessing and choosing 5 instances of a product with lowest price from multiple web domains advertising that product. User can also specify particular web domains for product price comparison that they considered best seller in terms of quality and price. This application is need of the time as in this digital world multiple online stores have been selling multiple products of same type and brand with different prices. Because of work load and human race users don't have enough time to access multiple sites to get a product with minimum price. The Figure 3 Shows the sample user interface of PennyWise.

Before moving towards results, legality issues must be taken in consideration while implementing such type of web application [17]. This site is accessing data of different web domains to compare prices of similar products they are offering to users for buying purpose. The question that is it legal to access their data or not arises but as far legality issue is considered, this application is accessing only data that is on the web interface of e-commerce sites and according to fact that it is free for anyone who is accessing this site it is clear that no violation is done. Other important issue to be considered is updating of web domains by their administrators. The web domains that have been accessed are updated time by time by their administrators meantime, to access these updated changes, the strategy that has been used is accessing web page every time the user query for a particular product. So that each time updated page is used for further process. This strategy makes our application better by reducing memory requirement, as local database is not required to save data periodically. Along with this error in sense of product availability is also reduced i.e., if we have used database and a person searches a product which is out of stock in web domain but present in our database displays wrong results and generate ambiguity. When web page from each web domain for a transaction initiated by user is scraped, the scraped information may contain different items because of short query like "Samsung J5" will return mobiles as well as accessories relevant to this query. But our concern is to display only relevant mobiles on our web application to do this we do string matching i.e. if user selected a category of mobiles our system fetches details of each product scraped from web page and perform string matching on product titles to identify required products from each web domain. The overall system architecture is showed in fig. 4.

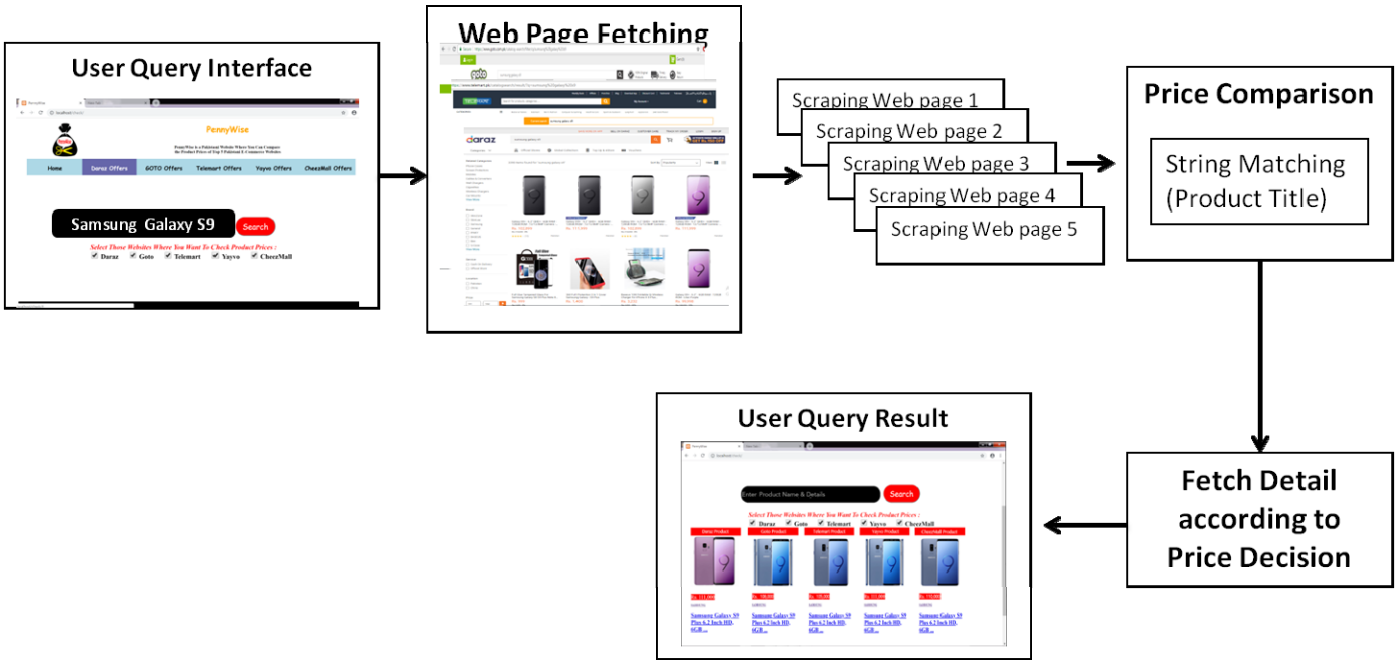

Fig. 4. System Architecture

TABLE: 1 Test Results

Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
Transaction	Result	Transaction	Result	Transaction	Result	Transaction	Result	Transaction	Result
T1	1	T1	1	T1	1	T1	1	T1	1
T2	1	T2	1	T2	0	T2	1	T2	1
T3	0	T3	1	T3	1	T3	1	T3	1
T4	1	T4	1	T4	1	T4	1	T4	1
T5	1	T5	1	T5	1	T5	1	T5	1

A. Information Retrieval Results

The details of any product including price, product specification and image etc. are extracted using web scraper. The web scraper gets the content and removes all unwanted data such as HTML tags and metadata except above mentioned details. The data retrieved is informative or not is checked by designing test cases. To know how accurate the data is scraped and displayed on interface, a test is conducted. For the test multiple queries are passed on to the system and identify how much data is retrieved from each web domain and how much of the shown data is close to user's wish.

1) Test Case

We have invited 5 subjects and requested them to make 5 queries to web application. After viewing result page, they are requested to give that query a score either 1 or 0. 1 is given to the page if it is according to the subject's wishes and 0 if it's not. The table 1 shows the transactions that have been intimated by subjects.

The overall success rate is based on accuracy. After evaluating subject's results accuracy is measured by using equation 1.

$$Accuracy = \frac{\sum_{k=1}^{n} X_k}{N} \quad (1)$$

Where N is the total number of transactions made by subjects, X is the successful transaction whose value is 1 for n transactions. The test results are showed in Table.1. Our system shows an accuracy of 93%.

IV. CONCLUSION

The internet is rich in term of massive data. With the passage of time, the data tomb is increasing. The need arises to fetch information using search engine which leads to more inconvenience in exact findings from the different sources over web and it is known that the contents available on web pages speaks a lot on massive topics[18]. Web scrapping techniques could help to reduce this problem.

Here, a system is developed to accumulate best e-commerce deals for clients from different web domains by using web scraping and price comparison. But People who are doing Scraping should take into account that they are not breaking any kind of law which could make them liable for any offence. In our scenario, we have access only data that is accessible by all viewers so it makes implementation of this application legally valid.

In future, we can develop mobile applications to facilitate mobile users.

REFERENCES

[1] V. Bhagwan and T. Grandison, "Injection," in *2009 IEEE International Conference on Web Services Deactivation*, 2009, pp. 2–3.

[2] H. Xuqldzdwl, G. Dndnrp, and D. F. Lg, "RQ H & RPPHUFH : HEVLWHV," pp. 5–8.

[3] F. Aulia and W. Dhewanto, "Formulation of E-Commerce Website Development Plan Using Multidimensional Approach for Web Evaluation," in *Procedia - Social and Behavioral Sciences*, 2014, vol. 115, no. Iicies 2013, pp. 361–372.

[4] S. Jie, S. Peiji, and F. Jiaming, "A Model for Adoption of Online Shopping : A Perceived Characteristics of Web as a Shopping Channel View," 2007, no. 2001, pp. 2001–2005.

[5] D. K. Mahto and L. Singh, "A Dive into Web Scraper World," pp. 689–693, 2016.

[6] S. Upadhyay, V. Pant, and S. Bhasin, "Articulating the Construction of a Web Scraper for Massive Data Extraction."

[7] L. Junjoewong, S. Sangnapachai, and T. Sunetnanta, "ProCircle : A promotion platform using crowdsourcing and web data scraping technique," *2018 Seventh ICT Int. Student Proj. Conf.*, pp. 1–5, 2018.

[8] L. R. Julian and F. Natalia, "THE USE OF WEB SCRAPING IN COMPUTER PARTS AND ASSEMBLY PRICE COMPARISON."

[9] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering- based approach to web advertising," vol. 2, no. 1, pp. 44–54, 2013.

[10] W. Scrapping and S. Annotation, "2011 International Conference on Computational Intelligence and Communication Systems," 2011.

[11] K. Sundaramoorthy, "NEWSONE- AN AGGREGATION SYSTEM FOR NEWS USING WEB SCRAPING METHOD," 2017.

[12] A. Saha, A. Singh, A. Kumar, and P. Kanjani, "Efficient and optimized use of data - Shrimp Browser," pp. 1135–1138, 2016.

[13] H. Lo, M. Reboiro-jato, F. Fdez-riverola, and D. Glez-pen, "Web scraping technologies in an API world," vol. 15, no. 5, pp. 788–797, 2013.

[14] C. Zheng, G. He, and Z. Peng, "A Study of Web Information Extraction Technology Based on Beautiful Soup," vol. 10, no. 6, pp. 381–387, 2015.

[15] N. Uppal and V. Chopra, "Design and Implementation in Selenium IDE with Web," vol. 46, no. 12, pp. 8–11, 2012.

[16] M. Monier and M. M. El-mahdy, "Evaluation of automated web testing tools," vol. 4, no. 5, pp. 405–408, 2015.

[17] A. J. Park and H. H. Tsang, "Phishing Website Detection Framework Through Web Scraping and Data Mining," pp. 680–684, 2017.

[18] M. Ying and Y. Hsu, "A Commodity Search System for Online Shopping Based on Ontology and Web Mining," 2003.