



Price Comparison of Computer Parts Using Web Scraping

Sanjay Kadam¹, Sanika Shinde², Akash Sharma³, Sonali Mali⁴Assistant Professor¹, BE Student^{2,3,4}

Department of Information Technology

Bharati Vidyapeeth College of Engineering, Navi Mumbai, India

Abstract:

Initially PCs were utilized just as a device to play out a few figuring's. These days PCs have a great deal of capacities to help individuals completing their assignments in practically every part of human life. In spite of PCs have a great deal of different capacities; clients likewise require some extraordinary particulars for every PC so they can do their assignments as per PC functionalities. Web scraping, the extraction of formatted data from web pages on the internet, have been developed in the private sector for business purposes, but it offers substantial benefits to those searching for grey literature. By building and sharing protocols that obtain search results and other data from web pages, those looking for computer related accessories can drastically increase their clarity and resource efficiency. Along these lines, this application is manufactured with a reason to prescribe an answer for its clients in gathering PCs that suit their requirements. This application additionally has cost examination highlight in view of information sources recovered from five PC shops. So the clients can limit the expenses of obtaining PC parts and the PC less demanding. This examination highlight depends on an essential shopper's guideline which is fundamentally purchasers need to buy things at the better cost as well as expecting the best quality as could be allowed.

I. INTRODUCTION

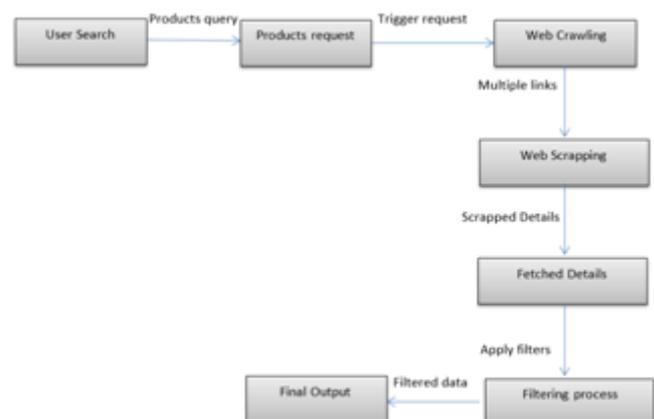
In the current era of online business, e-commerce have become a huge market for the people to buy goods online. Increasing use of smart devices and other mediums has paved the way for users to buy products almost from anywhere. This has increased involvement of online buyers evolving e-commerce business. These large numbers of e-commerce websites put users in turmoil to search and choose to buy a single product from multiple e-commerce websites. Through web scraping services unstructured data are converted into structured data which can be stored and verified in a centralized data bank. The aim is to collect, store and analyze data. The data analysis is very much needed in a society to extract any information and transforming it into a format helpful to understand. Thus, web scraping services have a direct influence on the outcome which is needed from the data collection. Web data extraction is the process of transforming the useful content on websites into valuable business assets. There are several web extracting software that has emerged in the market which helps to address this problem. The software helps in extracting structured content from a web page and exposes the required services as APIs and makes it useable for further processing.

II. LITERATURE REVIEW

The usage of web scraping as a strategy for information recovering that the creator utilized as a part of this examination has been directed before by different scientists who did other inquires about with various purposes and questions. Eloisa Vargiu and Mirko Urru (2012) beforehand have used web scratching in an electronic publicizing application with a reason to find if the promotions that will be posted in a website page is appropriate to the substance of the site page itself. Another examination was led by Maxim Bakaev and Tatiana Adveenko. They used web scratching in an examination about information recovery of occupation candidates, hopefuls who will be utilized by the administration

of work advertise in Novosibirsk, Russia. By utilizing this application, the association of work in Novosibirsk can break down the prerequisites of work that any organization require in that city with the similarity of work that related with the prerequisites. Other research that has executed web scratching was directed by three undergrads from Universitas Bina Dharma, Ahmat Josi, Leon Andretti Abdilah, and Suryayusra. The foundation of this examination was to facilitate the looking of logical articles required by the clients. They directed an examination to fabricate an application that can give different logical articles by utilizing web scratching. There is likewise an exploration that was using web scratching to break down the climate condition in Canada by Charmaine Bonifacio et al (2014). The application that was come about because of this examine was named Canadian Climate Data Scraping Tool (CCDST) where the information of the climate were straightforwardly acquired from the association of climate conjectures in Canada.

III. TECHNOLOGY USED



a. Search Product

User has to login to the application for comparing price of the desired computer accessories. As the user login, he/she can

fire the query. This query is hit on particular websites and also stored in database for requesting price details.

b. Web crawling

Web crawler is one of the main components of the Project. Since the product is price comparison engine, the first thing that is required is to collect large amount of data in terms of products from different websites. Manually, the collection of such large amount of data is not possible. So the best way to obtain these data is to create a web crawler also known as spider. For crawler to be more effective, it is necessary that the crawler is efficient, concurrent and multi-threaded. For crawler to be multi-threaded, it is important that the synchronization among the threads is maintained. So use of blocking queue came into picture. The main purpose of the crawler is to crawl different websites and to fetch the URLs of the products from these websites. Every website can be considered as a graph consisting of several nodes (Links or URLs). The crawler must pass through all these nodes and fetch these nodes. Once it has fetched the node, that node must be kept in a set of visited nodes so that no two same URLs are fetched. Threads that are created in the thread-pool must be limited so that they do not eat up the entire memory. And each thread that's been started has to be terminated. The Coordinating thread allocates the crawl job to the processing threads. These processing threads fetch the URLs and returns to the Coordinating threads. Thus the fetched URLs that we have in the set visited nodes are given to the scraper for scraping purpose after filtering.

c. Filtering Data

While filtering data, important task is to navigate to the destination once the crawler reaches the correct page and matches up with the products. It fetches information from websites so as to check for updates. If updates are available crawlers (i.e. filtering) carries those updates and makes necessary changes in the database. Also desired values are fetched through this filtering.

d. Web scrapping

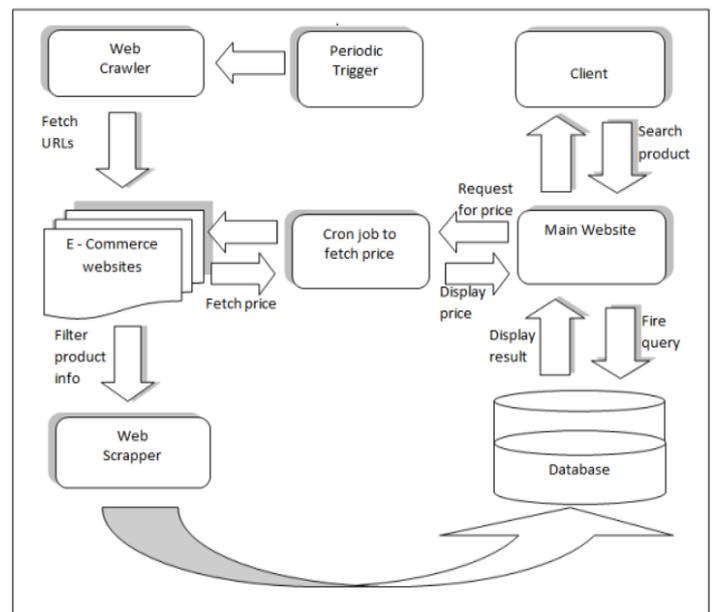
Web scraping can be defined as a process of extracting HTML data from the URLs and then using this data for personal purposes. Once we have the fetched URLs with us then the job is to get the information that is abstracted within the URL. For example the mentioned URL contains information such as the name of the product and the price and other related information on the link. This information is to be extracted for the purpose of comparison. The scrapper scrapes the information on this page on the basis of the tags in which the element is present. In this way the information can be extracted that are abstracted within the URLs. The extracted information is then stored in the database in the unstructured format. Once we have the fetched URLs in the visited set, then all the links in the visited set are to be scrapped. Once we have all the scraped information, we will store it in the database. Once we have data in the database, then the use of tool elastic search is being made for indexing and searching purpose. Indexing creates index of data through which searching becomes extremely easy. Once we have indexed data, then we are querying the database to get the duplicate products. Hence this is the implementation details in brief. Thus we get the final output of prices from various databases.

IV. PROPOSED SYSTEM

Web scrapping is a technique that is used to extract information in the human readable format and display it on destination terminal. But before scrapping the output, Web

Crawlers are responsible to navigate to the destination once the crawler reaches the correct page and matches up with the products, scrapping process starts. Crawler periodically fetches information from various websites so as to check for updates. If updates are available crawlers carries those updates and makes necessary changes in the database. When a user is comparing products on a website, the probability to buy is higher and the user is closer to the final sale. For the user, it is a matter of finding the best price. A simple comparison is an important tool for the user. Price comparison sites do other tasks - aggregate a large set of products across categories, aggregate deals in one place, aggregate coupons, build the widest catalogue for the user to browse through, trigger price changes, deep link to marketplaces etc. The best place to track value offered by such a site is to monitor it in infancy. Price comparison is the cheapest form of traffic for e-Commerce sites across the world. This is the single most important reason why their economic relevance is guaranteed for the foreseeable future.

V. SYSTEM ARCHITECTURE



Web scraping technology offers a wide range of options and may serve different objectives. The minimum requirement of a web crawler is to automate the usually manual work of collecting price quotes and article information from websites. The maximum requirement of a web crawler would be to explore price data sources previously unavailable and to provide a census of all price information available on the internet. The decisions made to set up web crawling for price statistics have important methodological and organizational impacts. In general, any price collection procedure with web crawlers will be composed of at least two steps: data extraction from website and the import of the extracted and validated price data to a data base. Price collection will be followed by cleaning and editing the data and a matching period. Working of the proposed system is as follows: The backend system consists of two important techniques web crawling and web scrapping. Web scraping or crawling is the process of obtaining data from a third party website by downloading and parsing the HTML code to extract the data you are interested in. Since every website does not provide a clean API, or an API at all, web scraping can be the only solution when it comes to extracting information from a

website. Many companies use it to obtain knowledge concerning competitor prices, news aggregation, mass email collect. Web scrapping is a technique that is used to extract information in the human readable format and display it on destination terminal. But before scrapping the output, Web Crawlers are responsible to navigate to the destination once the crawler reaches the correct page and matches up with the products, scrapping process starts. Crawler periodically fetches information from e-commerce websites so as to check for updates. If updates are available crawlers carries those updates and makes necessary changes in the database. Web scrapping essentially consists of two tasks: first is to load the desired web page and second is to parse HTML information of the page to locate intended information. Required results are retrieved and displayed on Main website. The client can then compare prices of products that are available on e-commerce websites. As soon as client selects on best deal according to him, he will be redirected to the original e-commerce website .Another feature provided is, Clients can compare products that belong to same category so as to differentiate specifications and choose accordingly.

VI. CONCLUSION

In light of the outcomes from client acknowledgment test which is led by the creator to ten respondents previously the application configuration is begun, the application has been prevailing with regards to noting the issues out of sight part and the detailing of the issues of this exploration where the outcome is an online use of cost correlation of PC parts and PC get together. This application will help the clients to get the proposal on the off chance that they need to assemble a PC. This application likewise enable the client to get the most reduced cost on the off chance that they need to purchase PC segments or assemble a pc.

VII. REFERENCES

- [1]. Jos´e Ignacio Fern´andez-Villamor, JacoboBlasco-Garc´ia, Carlos ´A. Iglesias, Mercedes GarijoDepartamento de Ingner´ia deSistemasTelem´aticos, Universidad Polit´ecnica de Madrid, Spain jifv@dit.upm.es, j.blasco@alumnos.upm.es, cif@dit.upm.es, mga@dit.upm.es
- [2]. Web Data Extraction, Applications and Techniques: A SurveyEmilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd
- [3]. WebSelF: A Web Scraping Framework Jakob Thomsen1, Erik Ernst1, Claus Brabrand2, and Michael Schwartzbach
- [4]. Exploiting web scraping in a collaborative filtering- based approach to web advertising.
- [5]. Faustina Johnson and Santosh Kumar Gupta.Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
- [6]. Cohen and Fan. Learning page-independent heuristics for extracting data from web pages. CN, 31(11-16), 1999.