Real Time Product Analysis using Data Mining

Jawahire Nakash*, Shaikh Anas, Siddiqi Muzammil Ahmad, Ansari Mohd. Azam, Prof Tabrez Khan

Abstract— **E-commerce is globally increasing business with increasing revenues every year manifold times. This is simple indication of more people moving online for shopping. They have developed many strategies by carefully analyzing the behavior of customers and overcoming the risk involved in online transactions to attract more business and participation from people. The Real Time Product analysis using data mining enables the buyers to compare products from different E-commerce websites, thus facilitating them to purchase the product at the best deal. To obtain the best deals from different E-commerce websites, a web crawler is used to crawl on different E-commerce websites and fetch the URLs of products. The scrapper scrapes the details abstracted within the URLs and stores it in the database. Then comparison among products of different E-commerce websites is made by using techniques such as inverted indexing. This way the paper aims to provide a solution which grants power in the hands of the users to purchase genuine products at genuine deal and saving user's time, money and efforts.**

***Index Terms*—** Business Analytics, Comparison Engine, Data Mining, Elastic-Search, JSoup, MongoDB, Product Analysis , Scrapper, Web Crawler.

I. INTRODUCTION

A large number of people nowadays gives the priority to on-line shopping over the Traditional Shopping, for that they are using smart devices such as tablets, mobile phones, laptop and desktop to access E-Commerce websites through the Internet. In addition, they want to get their desire product in best price. In order to get desire product in minimum price they survey or searches the number of E-commerce sites. To address these challenges, several weird agents-based e-commerce systems and add-on have been proposed. But all those sites and technology do not satisfy the users demand due to restriction, limitation of all these technology limited to its range of domain. The Real time product analysis overcomes these issues. This technology satisfied the user demand. Hereby using this technology user can get their desire product in minimum price apart from this features this technology recommends best to buy product to user over the numbers of E-commerce website.

E-commerce websites have a certain model and standards that are followed in the industry. In Real Time Product Analysis using Data Mining, Intelligent agent is used to crawl through to different websites to fetch URLs of different products. The intelligent agent is a web crawler or a Shop-Bot that is an automated program that fetches the URLs from different E-commerce websites. Most E-commerce websites only provide the products that are available with them at a particular rate. In most cases, users before purchasing the products on-line, they need to visit different E-commerce websites to find the particular products at the cheapest price. Real Time product Analysis using Data Mining solves this problem of user by providing user the products from different E-commerce websites at one place with different prices and schemes and offers that are offered by different E-commerce firms. This will provide the user privilege to choose products from different E-commerce websites which they consider is the best for them. The main elements of this technology are as follows:

A) Web Crawler B) Scraper

A) WEB CRAWLER:

Web crawler is one of the main component 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 E-commerce websites. Manually ,the collection of such large amount of data was not possible. So the best way to get 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 are maintained. So use of blocking queue came into picture. The main purpose of the crawler is to crawl different E-commerce websites and to fetch the URLs of the products from these websites. Every E-commerce website can be considered as a graph consisting of several nodes(Links or URLs)as shown in figure 1. The crawler must traverse to 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 distributes the crawl job to the processing threads. These processing threads fetches 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.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

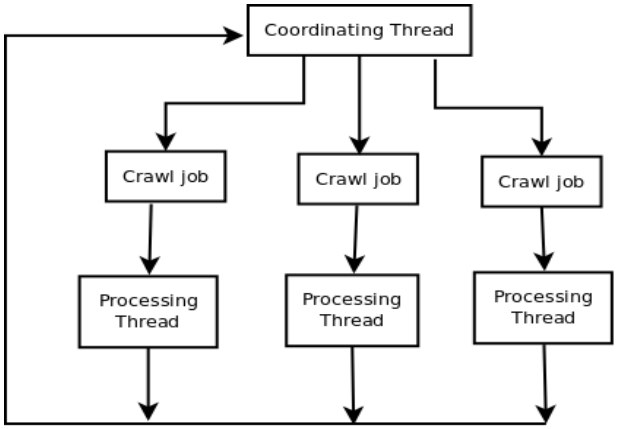


Figure 1. Flow Control of Web Crawler

B) SCRAPER

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[9] contains information such as the name of the product and the price and other related information on the link. These information are 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. 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.

II. LITERATURE REVIEW

Web Crawler/scrapers were addressed for extracting URLs from different E-commerce websites. In paper [1] focus was on product comparisons on behalf of humans. Also the focus was on implementing the architecture for online services by searching products from online websites and comparing the product amongst different websites and getting the cheapest price available on that product. The basic purpose was to search the product in cheapest price.

In Real Time Product Analysis using Data Mining, the intention is to provide the customer a user experience that allows the customer to view and compare prices of a particular product from different websites and purchase the product which he/she finds suitable for him/her. This tends to reduce Time and effort put by customer providing customer ease and satisfactory results. It also tends to save customer from predatory pricing strategies imposed by different E-commerce websites. In 2009 the functionality and performance of online shopping bots for E-commerce was investigated [2]. Eventually Real Time Product Analysis saves customers valuable time, efforts and money.

The technology being used is JSP& Servlet, Jsoup and MongoDB for backend processing while Html and CSS for front end. Jsoup is a specialized tool developed by MIT written in java language for URL extraction. For storing URLS and products, Nosql MongoDB is being used. Jsp and Servlet are use to develop the backbone structure of the comparative website. The hurdles and obstacles that need to be overcome: WebCrawler/scraper is the asset of the comparison website. Therefore, the crawler has to be very efficient in terms of fetching URls from different E-commerce websites in minimum time. This means, the crawler has to be very fast and efficient. Another Hurdle is data Storage that has to be overcome since there will be huge data present in different E-commerce websites with different naming standards that has to be taken into care. The procedure use is as follows:

- The crawler will continuously run in the background, fetching ULRs from different E-commerce websites.
- The fetched URls will be stored in the MongoDB database.
- The user needs to enter the query regarding particular product which he/she wants to purchase in the search bar.

As soon as the query is entered by the user, it triggers local searching algorithms to query the database to bring the required results. The crawler will bring periodic updates that are made by the original E-commerce websites and will update the local database. The customer compares different products based on prices and schemes available on that particular product given by different E-commerce websites. As soon as the customer clicks on the buy button on the product suitable according to him, this triggers the customer to original E-commerce website from where customer can do original purchasing of the product [3].

Thus comparison website is only acting as a mediator or an agent providing customer ease, reducing the effort and saving the money.

III. SYSTEM ARCHITECTURE.

E-commerce application have a certain model and standards that is followed in the industry. In Real Time product analysis using Data Mining, Intelligent agent is used to crawl through to different websites to fetch URL's of different products. The intelligent agent is a web crawler or a ShopBot that is an automated program that fetches the URL's from different E-commerce applications. Most E-commerce applications

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

only provide the products that are available with them at a particular rate. In most cases, users before purchasing the products online, they need to go different E-commerce applications to find the particular products at the cheapest price.

Real Time product Analysis using Data Mining solves this problem of user by providing the user Products from different E-commerce applications at one place with different prices and schemes and offers that are offered by different E-commerce firms. This will provide the user privilege to choose products from different E-commerce applications which they consider is the best for them. For this, we make use of machine Learning algorithms that keep Track of the user's behavior and searching patterns.

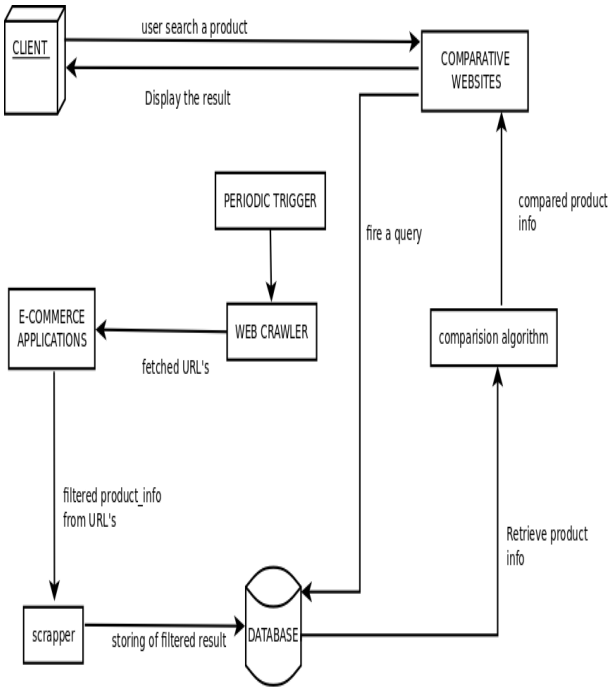


Figure 2. System Architecture of Real Time Product Analysis

IV. WORKING :

Figure 2 describes the system architecture and its working procedure. An Intelligent agent is a web crawler running on the back-end of the website whereas front end technology provides a graphical user interface (GUI) for the users to communicate with the system. The explanation of the architecture is as follows: The Web-Crawler visits different E-commerce websites and fetch URLs from different E-commerce websites. The Filter performs filtration so as to remove useless URLs. Then the filtered URLs are stored in the local database. The database used is MongoDB. Web crawler periodically fetches the data from different E-commerce websites and if updates are available, then web crawler carries the updates and updates the local database. Whenever client searches for the products in the search bar of the comparative website, the local database is queried so as to retrieve the required results. The user can then compare products based on prices from different E-commerce websites. When user selects the best deal according to him and click on the buy button of the product then on clicking on the buy button, it triggers the user to original website to purchase the product.

V. IMPLEMENTATION TECHNIQUES

In implementation of this project Real Time Product Analysis using data mining, the first step is to develop a web crawler which we are doing so by making use of java and jsoup. Since every website can be considered as a graph therefore in development of a WebCrawler, the most important technique is to make use of a linked blocking queue to traverse the graph ,that is to different links(nodes) of the website and to visit each of the link and store it in the visited set. Since the crawler is multithreaded , the main thread passes the task to the worker threads that visits different urls and store it in the visited set. In this way no two same urls are fetched .The linked Blocking Queue grants the flexibility for the threads to simultaneously offers links to the queue.

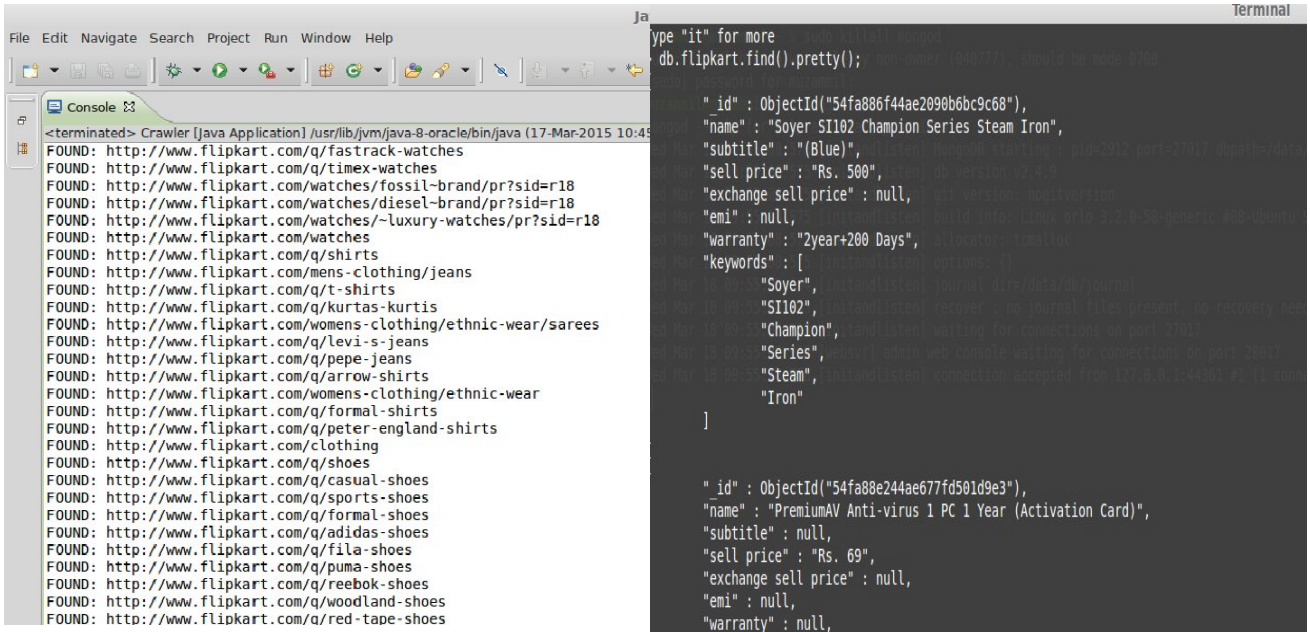
Once we have the fetched urls in the visited set, then all the links in the visited set are to be scraped. In scraping , by making use of Jsoup.connect(url).get we can get the information from the <img> tag, <h1> tag , tag etc. Once we have all the scraped information, we are storing it in the MongoDB 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.

VI. RESULT

We are comparing products of two E-commerce website. The webcrawler fetched Urls from these E-commerce websites. Example ,

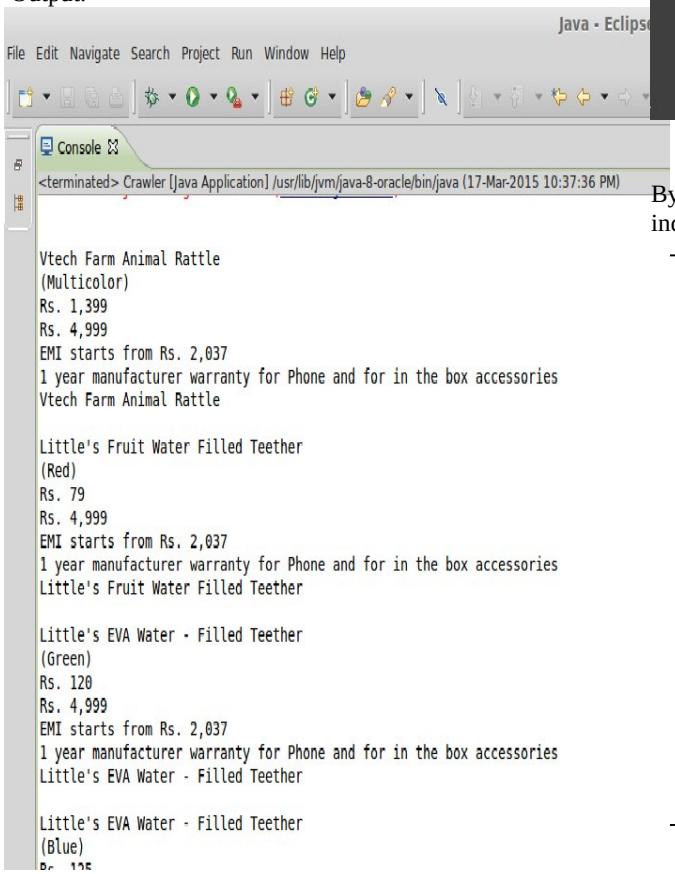
Below picture shows the urls that we are fetching from different E-commerce websites.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET



From these urls , we scrape the details of the product.

Output:



Then these scrape details we are adding into the database i.e. MongoDB.

By making use of tool elastic search , we are performing indexing on the product .

{
“productName”=Samsung galaxy grand 2
“sellerName”=gizmogear
“sellingPrice”=Rs 14,500
“EMI”=Rs 704
“Devilivery”=+ Rs100
}

{
“productName”=Samsung galaxy grand 2
“sellerName”=delhideals
“sellingPrice”=Rs 14,140
“EMI”=Rs 700
“Devilivery”=+ Rs100
}

Once we have the products all information, we can compare the products.
One Simple Elegant query is used to compare different products from the database:

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

```
String query = QueryBuilders.termsQuery("message",
Arrays.asList("trying", "out")).toString();
      SearchResponse searchResponse =
client.prepareSearch(indexName).setTypes(indexType)

.setSearchType(SearchType.QUERY_THEN_FETCH)
            .setFrom(0)
            .setSize(50)
            .setQuery(query).execute().actionGet();

System.out.println(searchResponse.getHits().totalHits());
      for(SearchHit hit:
searchResponse.getHits().getHits()) {
            System.out.println("***********");
            System.out.println(hit.getId());
            System.out.println(hit.getSourceAsString());
            System.out.println(hit.getScore());
```

This query will compare the products based on the relevance.

```
Output:
Samsung galaxy grand 2 from gizmogear
Rs. 14,500
Selling Price
EMI starts from Rs. 704 .
+ Rs 100 Delivery


Samung galaxy grand 2 from delhideals
Rs. 14,140
Selling Price
EMI starts from Rs. 700.
+ Rs 150 Delivery
```

VII. FUTURE SCOPE

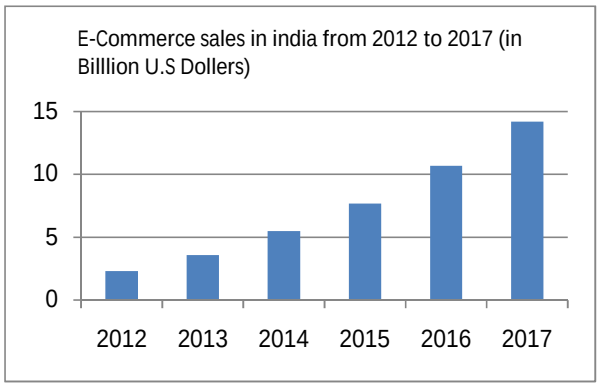
In the future, the product aims to include several features like suggesting the user about the expected future changes in prices on products on the basis of business analytics that can predict the situation and demand of the products in the market.

VIII. CONCLUSION

The Real Time product Analysis using Data mining is a price comparison engine that aims to facilitate the buyers to compare products from different E-commerce websites and purchase the product at the cheapest price with best deal. This way, the buyer has more power in his/her hands and can make better decisions. Thus this paper saves buyers efforts, time and money. It also helps the different E-commerce applications to boost their business by providing them a platform to compete and do business in a more reasonable manner.

Appendix

RETAIL E-COMMERCE SALES IN INDIA FROM 2012 TO 2017(IN BILLION U.S. DOLLARS)



ACKNOWLEDGMENT

We take this opportunity to express our deepest gratitude and appreciation to all those who have helped us directly or indirectly towards the successful completion of this Paper.

Foremost, we sincerely express our deep sense of gratitude to our guide **Prof. Khan Tabrez** and our co-guide **Prof. Farhaan Mir** for their advice, constant support, encouragement and valuable suggestions throughout the Paper completion helped us successfully complete the Paper. Without their continuous support and interest, this Paper would not have been the same as presented here.

Besides our guide, we would like to thank entire teaching and non-teaching staff in the Department of Computer Engineering for all their help during our tenure at AIKTC.

We also take this opportunity to thank whole-heartedly Honorable Director **Dr. Abdul Razak Honnutagi** and our HOD **Prof. Tabrez Khan** who has imparted valuable teaching and guidance that has inspired us to attain new goals.

REFERENCES

[1]. Stigler, G. 'The Economics of Information', Journal of Political Economy, Vol. 69, pp.213- 225, Iss. Jan-Feb, (1961).
[2]. Rehman S.U. 'Smart agent for automated E-commerce', Sch of inf Syst ,Comput. & Math., Brunel Univ, Uxbridge, UK IEEE pp.124-128, 7-10 Nov.2011.
[3]. Serenko and Hayes J 'Investigating the functionality and performance of online shopping bots for E-commerce', Electronic Business 8(1): pp.1-15,(2009).
[4]. Seddin K,Serenko and Hayes J ,'Online shopping bots for E-commerce', Int. J. Electronic Business, pp. 556-589, (2007)
[5]. Clark, D. 'Shopbots become agents for business change', IEEE Computer, Vol. 33, Issue 2, 2000.
[6]. http://www.postonline.co.uk/post/analysis/2319717/ the-rise-of-price-comparison-sites- in- south-east-asia
[7]. http://e27.co/former-lazada-thailand-ceo-now-working-on-

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

financial-products- comparison site/
[8]. http://www.consumerfutures.org.uk/
[9]. http://www.snapdeal.com/product/htc-desire-526-g/ 681509924248

AUTHORS PROFILE

Mr. Jawahire Nakash pursuing B.E in Computer Engineering from Anjuman-I-Islam's Kalsekar Technical Campus affiliated to Mumbai University, India.

Mr. Shaikh Anas pursuing B.E in Computer Engineering from Anjuman-I-Islam's Kalsekar Technical Campus affiliated to Mumbai University, India.

Mr. Siddiqi Muzammil Ahmad pursuing B.E in Computer Engineering from Anjuman-I-Islam's Kalsekar Technical Campus affiliated to Mumbai University,India.

Mr. Ansari Mohd. pursuing B.E in Computer Engineering from Anjuman-I-Islam's Kalsekar Technical Campus affiliated to Mumbai University, India.

Prof. Tabrez Khan is an Assistant Professor in department of Computer Engineering, Anjuman-I-Islam's Kalsekar Technical Campus, Navi Mumbai, affiliated to Mumbai University, India.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET