Using Unstructured Big Data and Web Content Mining for Price Spread Analytics

Syed Imtiyaz Hassan, Md. Omair Ahmad, M. Afshar Alam
Department of CSE, SEST, Jamia Hamdard (Deemed to be University), New Delhi, India.

ABSTRACT

The usage of online shopping portal has made the shopping easy in a sense that one can purchase a product without visiting the shop physically. In another way, it is also difficult to shop online as there are a number of similar product offered by different online shopping portal with difference in cost. So, for getting value for money one has to explore each portal and then need to compare the cost. It is a difficult, time consuming and sometimes confusing task. As the data generated from such searches are unstructured big data, a specialized techniques must be applied. This research is an attempt to make the life of online buyers easy by mining some valuable information that a customer generally looks for. By processing unstructured big data generated from online shopping portal searches, the proposed technique extract minimum (or maximum) cost product offered by various online shopping portal, product wise and date wise. Along with the minimum cost of various product searched, minimum cost of other related products have also been mined. For this purpose five data sets have been generated from searches of various online shopping portals. Hadoop and Map Reduce is used to process large amount of data set in parallel. Java and Eclipse IDE have been used for necessary coding and for execution environment. Tibco Spotfire is used for graphical visualization and data analytics.

KEYWORDS*:* Data Mining, Data Analytics, Big data, Web Content Mining, Hadoop Framework, Map Reduce, Tibco Spotfire

1. INTRODUCTION

Due to the growing usage of e-commerce applications and ease of shopping, the demand for online shopping is increasing day by day. This increased demand is producing large amount of data. The continuous addition in data size is raising difficulty while dealing with it [1]. The enterprises are bound to deal with big data that are up to many terabytes in size. As the size of these data are huge and are generally unstructured, processing of data within acceptable time limit, reducing the cost storage management and finding useful information is a big issue [2].

With time web portals are becoming much accepted for information retrieval, knowledge discovery and online shopping [3]. That is why business try to utilize web log for finding the patterns. These web logs [4, 5] are generally created and maintained by a web server that stores the activities performed by a user in a particular website. The various information stored in a web log may include person identity and the actions performed. Discovering such pattern from the web comes under the purview of web mining [6].

Whenever a customer search for a particular product while online shopping, (s) he gets same product in different online store with different cost. Together with the original product searched (s) he also gets an options of selecting a related product that other customer also viewed. Sometimes it rather confuses the customer. This research is intended to make customer comfortable while taking final decision. It also focus on managing of the most viewed and demanding product on the seller’s site. In short this research focus is web content mining of big data [7, 8] with a case of online shopping.

1.1. WEB CONTENT MINING

Web is a collection of documents. The web content is incredibly huge, flexible and dynamic. The WWW has grown in the large volume of traffic, size and complexity of Internet sites. It is difficult to spot the helpful, needed and desired information present in the web. The growing field of web mining aims for finding and extracting relevant information that is hidden in web connected knowledge, in particular in text documents. Data mining [9, 10] is that the idea of extraction valuable information from massive volume of information. Web mining is a specialized application of data mining to mine knowledge from web data using web content, web structure, and web usage data.

The information handled by a web site in general are comprise of content, structure and log data. The process of web mining based on these categories may be further divided as: web content mining, web structure mining and web usage mining. As the name suggests, web content mining deals with the presented data, the web structure mining [11] is all about extracting web site structure, and web usage mining [12] explores the usage characteristics by web site users. All these mining process extract information semi-structured nature, and hence requires various pre-processing and parsing techniques to be applied for better result. Some of the prominent are domain understanding, data pre-processing, data cleansing, pattern discovery, interpretation, and reporting. However the main focus of this research is web content mining.

1.2. BIG DATA

The quantity of data generated by web is massive [13]. The term big data has been used to describe such huge data that exceeds the process capability [6] of humans. It has been characterized by 5V [14]: Volume, Velocity, Variety, Variability and Veracity.

- Volume: The magnitude of knowledge is massive; generally in terabytes and petabytes. This huge volume of large size of knowledge makes it difficult to analyze using traditional methods.
- Velocity: Since the rate by which data is generated is fast, it shall be processed in restricted time frame. The conventional methods of data mining technique may not be applied in such cases.
- Variety: The sources from where big data is generated are very different in nature. So, big data methodology must be able to deal with all format of data; structured, semi-structured and unstructured.
- Variability: Inconsistency is an issue for big data that make it difficult to store, process and manage.
- Veracity: The quality of data can vary significantly that may affect accurate study of patterns.

1.3. PROBLEM STATEMENT AND OBJECTIVE

A typical online trading application needs to process thousands of customer's transactions per day that is also needed to be stored in its database. All such data transactions need to be clustered according to some chosen parameters for identification of meaningful patterns and inference. An example of such pattern is probability of buying of related product based on a purchase. As with rapid development of Internet, e-commerce where data are mostly unstructured or semi structured and huge [15], the older techniques like clustering [16] is not fast and efficient. In such situation finding relevant and useful data is difficult and time consuming.

To address the above problem, an appropriate web mining technique has been selected. Further raw data is converted into structured data. For this purpose machine learning [17-20] has been used to mine big data. The aim of present research is to consider the following issues while keeping in mind the growing size of online products data:

- To filter the product based on web data log such as frequent access paths, frequent seen products etc.
- To transform the data in structured format for meaning.
- To make e-commerce and online shopping selection item easy and more accurate.

2. REVIEW OF LITERATURES

A work has been published by [21] to study the problem of extracting knowledge records from the website managing huge data. To evaluate the tactic and concepts projected during this research, a large variety of experiments have been conducted to demonstrate their effectiveness. Further [22] proposed a methodology for extracting data from the Internet in knowledge acquisition. The basic idea of the research was to bound lexico-syntactic patterns and semantic relation. In [23] a research has been published on web content mining. The main focus was on managing web content between the server and the client. It also discussed the issue of logs. A very innovative work has been proposed by [24] that introduced a new way of identifying semantic correspondences by usage-based relations between the objects themselves.

In recent years various system has been proposed using web content mining. The paper [25] presents a model to form a reliable forecasting structure for the US presidential elections and US House race. A novel web service recommendation system by [26] has been proposed by integrating a user's probable QoS preferences and variety of user likeliness on web services. A system VAiRoma [27] has been proposed based on visual analytics approach to help consumers make sense of places, events, times and the relations between them based on a big pool of Wikipedia articles.

Further a work was proposed by [28] to keep track of active users and the television events. Based on this information the popularity of television event may be known. A software architecture has been developed by [29] to make and sustain a Genomic and Proteomic Knowledge Base (GPKB), which assimilates a number of the most pertinent sources of such discrete information.

As the work of mining unstructured big data is multi-disciplinary and exhaustive, it is very difficult to list of all the works that has been done in this direction. However, it has been attempted to mention some notable work the work in the domain of this research.

3. PROPOSED SOLUTION AND IMPLEMTATION

To mine the minimum and maximum cost of products from unstructured big data generated from online shopping portal searches following steps have been performed:

1. A search on various product has been performed by visiting various portals for few days.
2. The unstructured data generated from these searches has been converted into structured data.
3. Apache Hadoop [30], Map Reduce algorithm [31, 32], Java code [33] and Eclipse IDE [34] have been used for processing and for generating necessary outcome.
4. Finally, Tibco Spotfire [35], a data analytics tool has been used for visualization.

3.1. CONVERTING UNSTRUCTURED DATA INTO STRUCTURED DATA

As the data extracted from web sites are generally unstructured in nature, it is difficult to process. That is why an attempt has been made to convert the unstructured data into structured form. Machine learning technique is used to extract the raw data.

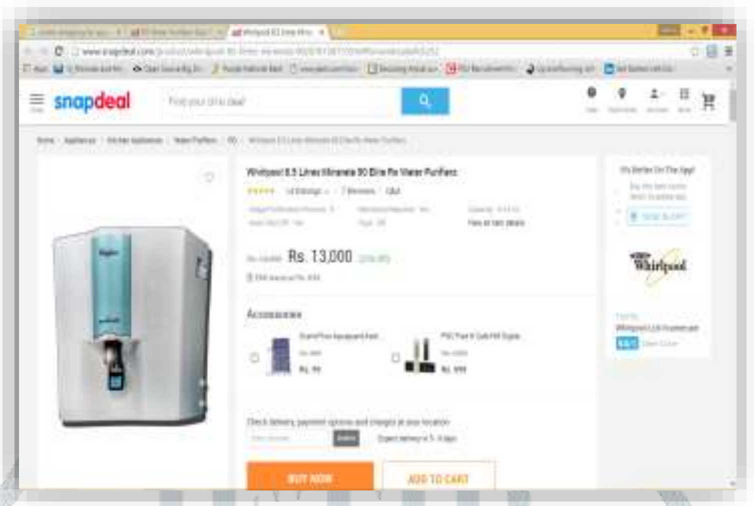
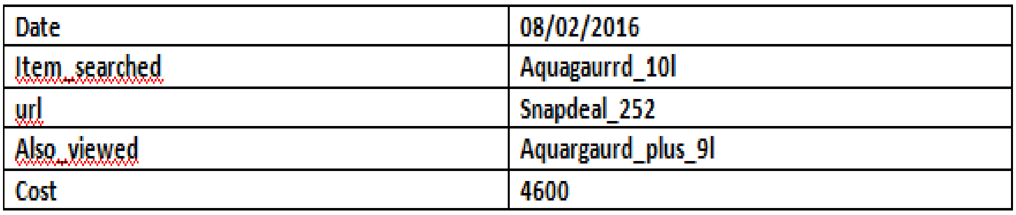


Figure 1. Sample instance of raw data after searched

For example a sample search depicted in figure 1 has been converted into structured data depicted in Table 1.

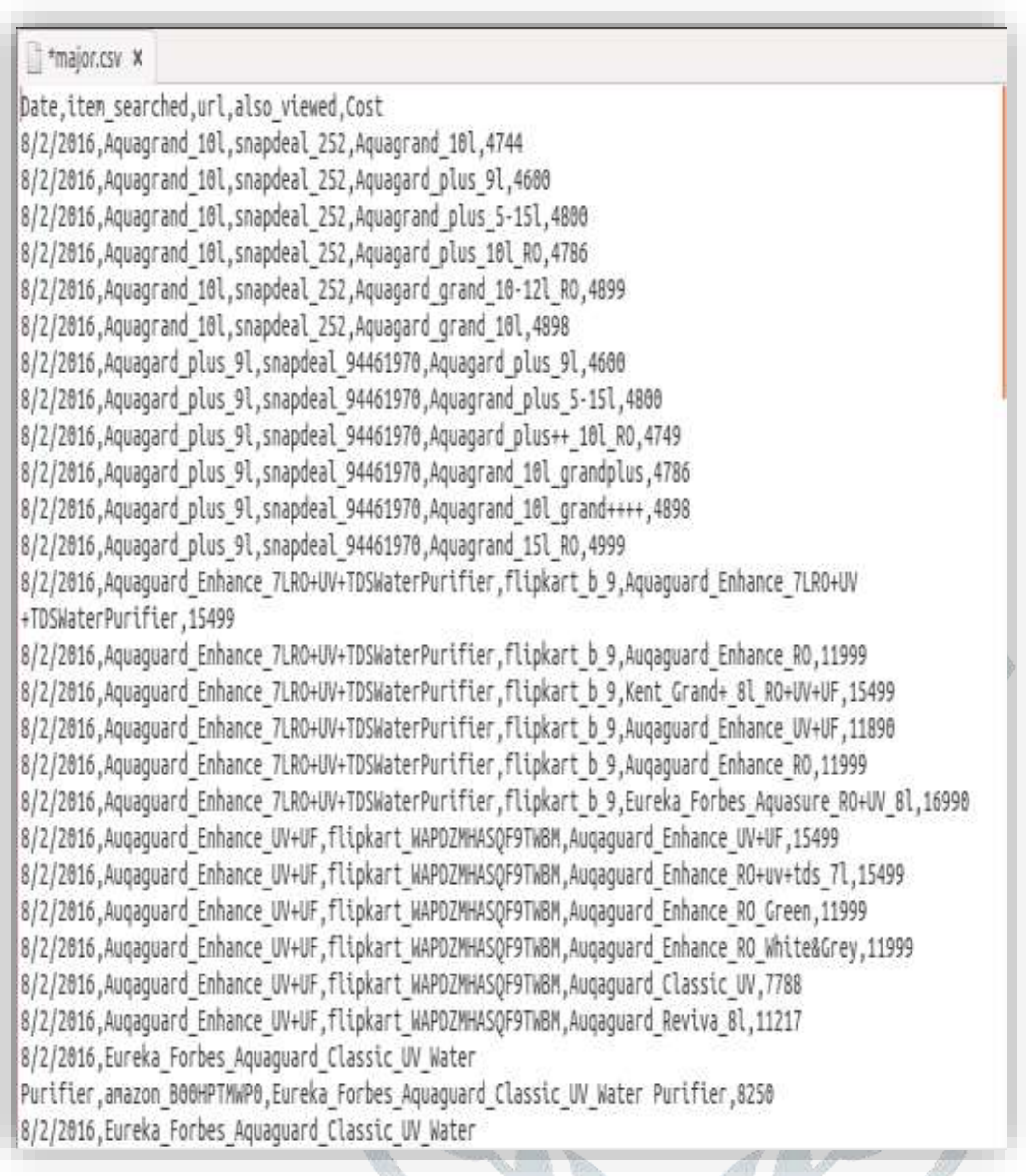


Date	08/02/2016
Item_searched	Aquagaurrd_10l
url	Snapdeal_252
Also_viewed	Aquargaurd_plus_9l
Cost	4600

Table 1. Structured data after conversion of raw data

3.2. DATA SETS USED

The data has been generated from various online shopping site such as Snapdeal, Flipkart, Amazon etc. Search of items were conducted for few days. This product search activity was recorded and stored into .csv file that is a form of structured data. A sample snapshot has been depicted in figure 2.



```
Date,item_searched,url,also_viewed,Cost
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagrand_10l,4744
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagard_plus_9l,4600
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagrand_plus_5-15l,4800
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagard_plus_10l_RO,4786
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagard_grand_10-12l_RO,4899
8/2/2016,Aquagrand_10l,snapdeal_252,Aquagard_grand_10l,4898
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagard_plus_9l,4600
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagrand_plus_5-15l,4800
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagard_plus++_10l_RO,4749
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagrand_10l_grandplus,4786
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagrand_10l_grand++++,4898
8/2/2016,Aquagard_plus_9l,snapdeal_94461970,Aquagrand_15l_RO,4999
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Aquaguard_Enhance_7LRO+UV
+TDSWaterPurifier,15499
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Auqaguard_Enhance_RO,11999
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Kent_Grand+_8l_RO+UV+UF,15499
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Auqaguard_Enhance_UV+UF,11890
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Auqaguard_Enhance_RO,11999
8/2/2016,Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier,flipkart_b_9,Eureka_Forbes_Aquasure_RO+UV_8l,16990
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Enhance_UV+UF,15499
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Enhance_RO+uv+tds_7l,15499
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Enhance_RO_Green,11999
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Enhance_RO_White&Grey,11999
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Classic_UV,7788
8/2/2016,Auqaguard_Enhance_UV+UF,flipkart_WAPDZMHASQF9TWBM,Auqaguard_Reviva_8l,11217
8/2/2016,Eureka_Forbes_Aquaguard_Classic_UV_Water
Purifier,amazon_B00HPTMWP0,Eureka_Forbes_Aquaguard_Classic_UV_Water Purifier,8250
8/2/2016,Eureka_Forbes_Aquaguard_Classic_UV_Water
```

Figure 2. Snapshot of the data Set used

Five different .csv files has been generated from activity log.

3.3. PROCESSING THE DATA WITH HADOOP AND MAP REDUCE

Each file is loaded into Hadoop environment through Hadoop Distributed File System abbreviated (HDFS). Map Reduce algorithm is used to process the data as it allows processing in parallel. It combines Map program and Reduce program. Map does filtering and organizing whereas Reduce accomplishes an output outline operation.

3.4. PROGRAMMING ENVIRONMENT USED

The proposed solution of the mentioned problem has used Ubuntu operating system. Hadoop and Map Reduce is used to process large amount of data set in parallel. Programming language Java is used for necessary coding for experiment. Eclipse IDE has been used for execution environment. Tibco Spot fire is used for graphical representation of the filtered data. It is a data visualization and analytics tool that helps to quickly uncover insights for better decision-making.

3.5. EXPERIMENTAL SEQUENCE OF EVENTS FOR HADOOP

The following sequence of events has been performed in the experiment for big data mining:

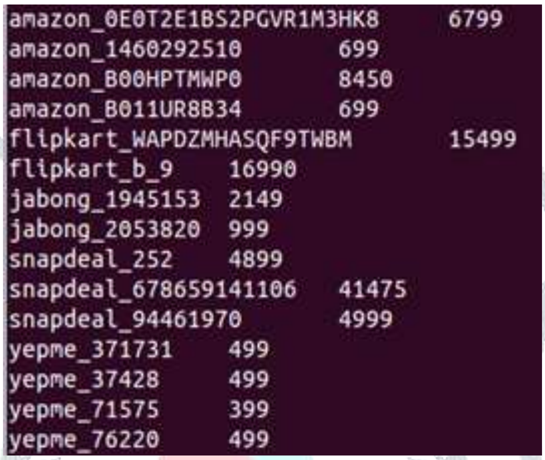
1. Load the data set in HDFS.

2. Write the Java code using Eclipse IDE
3. Make a jar file of the package to execute it.
4. Load the output/result in HDFS

The above process is repeated separately for all the five .csv files. Filtering is done on the basis of date of search, URL used, items searched and item viewed. Key value pair is generated for all keys. Each time the cost is evaluated. Finally cost range has been produced. Maximum and minimum cost of the searched products as well as the viewed products has been generated.

4. RESULTS AND DISCUSSION

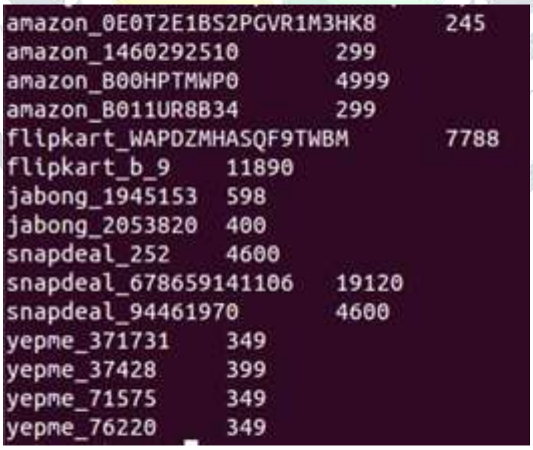
Whenever a customer search for a particular product, that product might be available in various online shopping portals. One has to compare the cost offered by different online shopping portal manually. Figure 3 shows the maximum cost of a product offered various portal.



```
amazon_0E0T2E1BS2PGVR1M3HK8     6799
amazon_1460292510       699
amazon_B00HPTMWP0       8450
amazon_B011UR8B34       699
flipkart_WAPDZMHASQF9TWBM       15499
flipkart_b_9    16990
jabong_1945153  2149
jabong_2053820  999
snapdeal_252    4899
snapdeal_678659141106   41475
snapdeal_94461970       4999
yepme_371731    499
yepme_37428     499
yepme_71575     399
yepme_76220     499
```

Figure 3. Maximum cost product offered by various online shopping portal

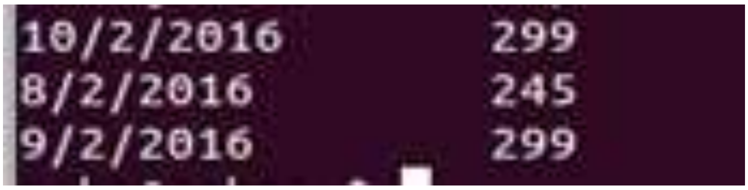
Similarly, Figure 4 depicts the minimum cost product offered by various online shopping portal. The mining has been performed on the basis of URL.



```
amazon_0E0T2E1BS2PGVR1M3HK8     245
amazon_1460292510       299
amazon_B00HPTMWP0       4999
amazon_B011UR8B34       299
flipkart_WAPDZMHASQF9TWBM       7788
flipkart_b_9    11890
jabong_1945153  598
jabong_2053820  400
snapdeal_252    4600
snapdeal_678659141106   19120
snapdeal_94461970       4600
yepme_371731    349
yepme_37428     399
yepme_71575     349
yepme_76220     349
```

Figure 4. Minimum cost product offered by various online shopping portal

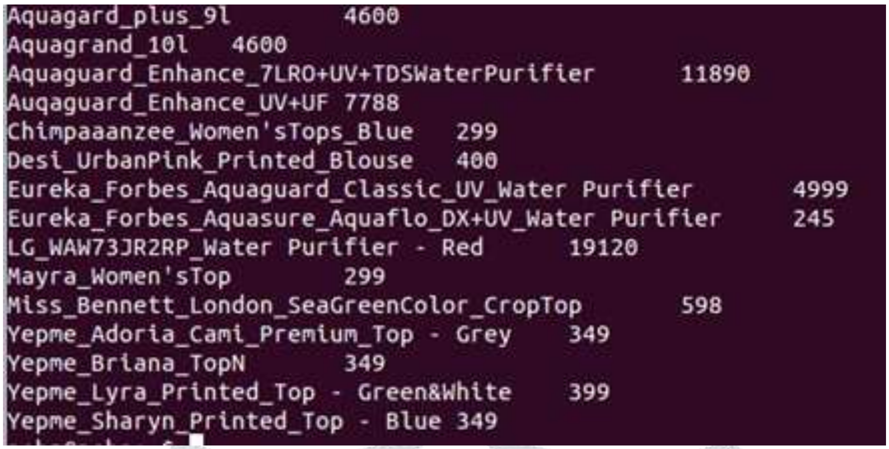
Sometimes, the cost of the product varies when searched in different dates. One can also mine the minimum (or maximum) cost of a product date wise (refer Figure 5).



```
10/2/2016       299
8/2/2016        245
9/2/2016        299
```

Figure 5. Minimum cost of a product date wise

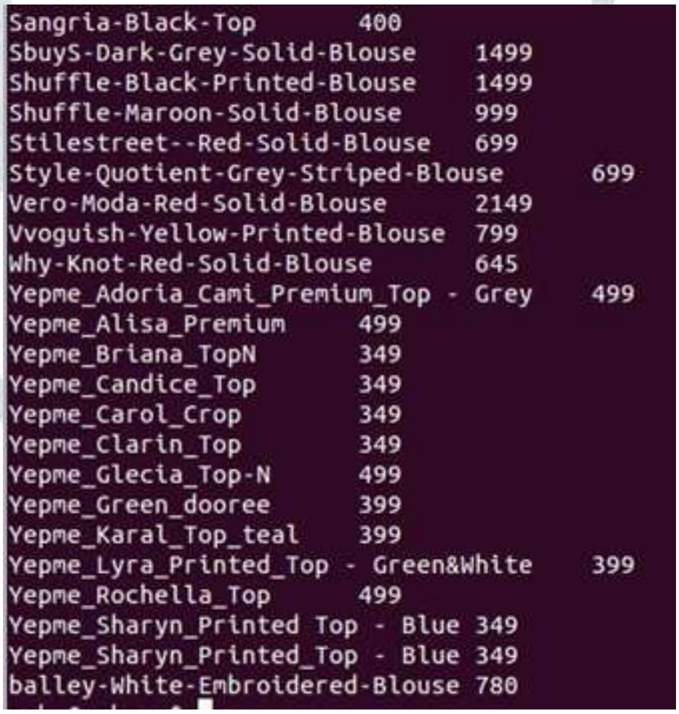
When a customer searches a product, (s) he searches various models of it. Figure 6 depicts the minimum cost of various models of a product of Aguaguard and women's top.



```
Aquagard_plus_9l        4600
Aquagrand_10l   4600
Aquaguard_Enhance_7LRO+UV+TDSWaterPurifier      11890
Auqaguard_Enhance_UV+UF 7788
Chimpaaanzee_Women'sTops_Blue   299
Desi_UrbanPink_Printed_Blouse   400
Eureka_Forbes_Aquaguard_Classic_UV_Water Purifier       4999
Eureka_Forbes_Aquasure_Aquaflo_DX+UV_Water Purifier     245
LG_WAW73JR2RP_Water Purifier - Red      19120
Mayra_Women'sTop        299
Miss_Bennett_London_SeaGreenColor_CropTop       598
Yepme_Adoria_Cami_Premium_Top - Grey    349
Yepme_Briana_TopN       349
Yepme_Lyra_Printed_Top - Green&White    399
Yepme_Sharyn_Printed_Top - Blue 349
```

Figure 6. Minimum cost of various product searched

There are a situation when a customer searches a product, shopping portal also suggests the link of some related product, that users clicks and explores. One can also mine the minimum (or maximum) cost of all these products. Figure 7 depicts the same situation.



```
Sangria-Black-Top       400
SbuyS-Dark-Grey-Solid-Blouse    1499
Shuffle-Black-Printed-Blouse    1499
Shuffle-Maroon-Solid-Blouse     999
Stilestreet--Red-Solid-Blouse   699
Style-Quotient-Grey-Striped-Blouse      699
Vero-Moda-Red-Solid-Blouse      2149
Vvoguish-Yellow-Printed-Blouse  799
Why-Knot-Red-Solid-Blouse       645
Yepme_Adoria_Cami_Premium_Top - Grey    499
Yepme_Alisa_Premium     499
Yepme_Briana_TopN       349
Yepme_Candice_Top       349
Yepme_Carol_Crop        349
Yepme_Clarin_Top        349
Yepme_Glecia_Top-N      499
Yepme_Green_dooree      399
Yepme_Karal_Top_teal    399
Yepme_Lyra_Printed_Top - Green&White    399
Yepme_Rochella_Top      499
Yepme_Sharyn_Printed Top - Blue 349
Yepme_Sharyn_Printed_Top - Blue 349
balley-White-Embroidered-Blouse 780
```

Figure 7. Minimum cost of various product searched along with related products

These related products is generally suggested under the heading *'also viewed'* or *'you may also like'*. It give customer a variety of similar options. Thus a better product is selected when number of option is increased and thus better decision can be made.

5. CONCLUSION

The main objective of the present research is to find the cost spread (the mining of minimum and maximum cost) of products from unstructured big data generated from online shopping portal searches. For this purpose five data sets has been created by searching various products on Snapdeal, Flipkart, Amazon and other online shopping portals. Hadoop and Map Reduce has been used to manage large date size and parallel processing. Finally, minimum and maximum cost of a product has been mined on various criteria.

REFERENCES

[1] R. Ranjan, L. Wang, A. Y. Zomaya, D. Georgakopoulos, X. H. Sun and G. Wang, "Recent advances in autonomic provisioning of big data applications on clouds," in IEEE Transactions on Cloud Computing, vol. 3, no. 2, pp. 101-104, April-June 1 2015. doi: 10.1109/TCC.2015.2437231

[2] Kun Wang, Jun Mi, Chenhan Xu, Qingquan Zhu, Lei Shu, and Der-Jiunn Deng. 2016. Real-Time Load Reduction in Multimedia Big Data for Mobile Internet. *ACM Trans. Multimedia Comput. Commun. Appl.* 12, 5s, Article 76 (October 2016), 20 pages. DOI: https://doi.org/10.1145/2990473

[3] Erhard Rahm. 2014. Discovering product counterfeits in online shops: A big data integration challenge. *J. Data and Information Quality* 5, 1-2, Article 3 (September 2014), 3 pages. DOI=http://dx.doi.org/10.1145/2629605

[4] Fedja Hadzic and Michael Hecker. 2011. Alternative Approach to Tree-Structured Web Log Representation and Mining. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01* (WI-IAT '11), Vol. 1. IEEE Computer Society, Washington, DC, USA, 235-242. DOI=http://dx.doi.org/10.1109/WI-IAT.2011.156

[5] R. Kanagasabai, A. Veeramani, H. Shangfeng, K. Sangaralingam and G. Manai, "Classification of massive mobile web log URLs for customer profiling & analytics," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1609-1614. doi: 10.1109/BigData.2016.7840771

[6] Xiao Fang and Olivia R. Liu Sheng. 2004. LinkSelector: A Web mining approach to hyperlink selection for Web portals. *ACM Trans. Internet Technol.* 4, 2 (May 2004), 209-237. DOI=http://dx.doi.org/10.1145/990301.990306

[7] N. Zulkarnain and M. Anshari, "Big data: Concept, applications, & challenges," *2016 International Conference on Information Management and Technology (ICIMTech)*, Bandung, Indonesia, 2016, pp. 307-310. doi: 10.1109/ICIMTech.2016.7930350

[8] Shailesh Singh and Syed Imtiyaz Hassan, "Detecting duplicates and near duplicates records in large datasets, *International Journal on Computer Science and Engineering (IJCSE),* vol. 9, no. 5 May 2017, ISSN : 0975-3397, pp. 178-185.

[9] C. Zhang, Q. Yuan and J. Han, "Bringing Semantics to Spatiotemporal Data Mining: Challenges, Methods, and Applications," *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, San Diego, CA, USA, 2017, pp. 1455-1458.
doi: 10.1109/ICDE.2017.210

[10] Jiuyong Li. 2017. Beyond Understanding and Prediction: Data Mining for Action. In *Proceedings of the 26th International Conference on World Wide Web Companion* (WWW '17 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1361-1361. DOI: https://doi.org/10.1145/3041021.3053407

[11] Hung-Yu Kao, Jan-Ming Ho and Ming-Syan Chen, "WISDOM: Web intrapage informative structure mining based on document object model," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 5, pp. 614-627, May 2005. doi: 10.1109/TKDE.2005.84

[12] O. Nasraoui, M. Soliman, E. Saka, A. Badia and R. Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 202-215, Feb. 2008. doi: 10.1109/TKDE.2007.190667

[13] Syed Imtiyaz Hassan, "Extracting the sentiment score of customer review from unstructured big data using Map Reduce algorithm*", International Journal of Database Theory and Application*, vol. 9, issue 12, Dec 2016, pp. 289-298, doi: 10.14257/ijdta.2016.9.12.26, ISSN: 2005-4270.

[14] Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. Development Policy Review, 34(1), 135–174. http://doi.org/10.1111/dpr.12142.

[15] V. Padmapriya, J. Amudhavel, V. Gowri, K. Lakshmipriya, S. Vinothini, and K. Prem Kumar. 2015. Demystifying Challenges, Opportunities and Issues of Big Data Frameworks. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (ICARCSET '15). ACM, New York, NY, USA, , Article 45 , 5 pages. DOI: http://dx.doi.org/10.1145/2743065.2743110

[16[Y. Jeon, J. Yoo, J. Lee and S. Yoon, "NC-Link: A New Linkage Method for Efficient Hierarchical Clustering of Large-Scale Data," in *IEEE Access*, vol. 5, no. , pp. 5594-5608, 2017. doi: 10.1109/ACCESS.2017.2690987

[17] Samir Kumar Singha and Syed Imtiyaz Hassan, "Enhancing the classification accuracy of noisy dataset by fusing correlation based feature selection with k-nearest neighbour", *Oriental Journal of Computer Science and Technology*, vol. 10, no. 2, June 2017, ISSN: 0974-6471, Online ISSN: 2320-8481.

[18] Baby Kahkeshan and Syed Imtiyaz Hassan, "Assessment of accuracy enhancement of back propagation algorithm by training the model using deep learning", *Oriental Journal of Computer Science and Technology*, vol. 10, no. 2, June 2017, ISSN: 0974-6471, Online ISSN: 2320-8481.

[19] Naba Suroor and Syed Imtiyaz Hassan, "Identifying the factors of modern day stress using machine learning*", International Journal of Engineering Science and Technology*, vol. 9, Issue 4, April 2017, pp. 229-234, e-ISSN: 0975–5462, p-ISSN: 2278–9510.

[20] Syed Imtiyaz Hassan, "Designing a flexible system for automatic detection of categorical student sentiment polarity using machine learning", *International Journal of u- and e- Service, Science and Technology*, vol. 10, no.3, Mar 2017, pp. 25-32, doi: 10.14257/ijunesst.2017.10.3.03, ISSN: 2005-4246.

[21] Zhang, Lakshmanan, and Zamar, "Extracting Relational Data from HTML Repositories", SIGKDD Explorations Volume 6, Issue 2, 2004.

[22] P. Cimiano, A. Hotho and S. Staab (2005) "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis", Journal of Artificial Intelligence Research, Volume 24, pages 305-339

[23] Dumitru Ciobanu, Claudia Elena Dinucă, WEB CONTENT MINING, Annals of the University of Petroani, Economics, 12(1), 2012, 85-92

[24] K. Niemann and M. Wolpers, "Creating Usage Context-Based Object Similarities to Boost Recommender Systems in Technology Enhanced Learning," in IEEE Transactions on Learning Technologies, vol. 8, no. 3, pp. 274-285, July-Sept. 1 2015. doi: 10.1109/TLT.2014.2379261

[25] Q. You, L. Cao, Y. Cong, X. Zhang and J. Luo, "A Multifaceted Approach to Social Multimedia-Based Prediction of Elections," in IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2271-2280, Dec. 2015. doi: 10.1109/TMM.2015.2487863

[26] G. Kang, M. Tang, J. Liu, X. (. Liu and B. Cao, "Diversifying Web Service Recommendation Results via Exploring Service Usage History," in IEEE Transactions on Services Computing, vol. 9, no. 4, pp. 566-579, July-Aug. 1 2016. doi: 10.1109/TSC.2015.2415807

[27] I. Cho, W. Dou, D. X. Wang, E. Sauda and W. Ribarsky, "VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 210-219, Jan. 31 2016. doi: 10.1109/TVCG.2015.2467971

[28] R. G. Pensa, M. L. Sapino, C. Schifanella and L. Vignaroli, "Leveraging Cross-Domain Social Media Analytics to Understand TV Topics Popularity," in IEEE Computational Intelligence Magazine, vol. 11, no. 3, pp. 10-21, Aug. 2016. doi: 10.1109/MCI.2016.2572518

[29] M. Masseroli, A. Canakoglu and S. Ceri, "Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 2, pp. 209-219, March-April 1 2016. doi: 10.1109/TCBB.2015.2453944

[30] Apache Hadoop haome page, http://hadoop.apache.org/

[31] Y. V. Lokeswari and Shomona Gracia Jacob. 2016. A Comparative study on Parallel Data Mining Algorithms using Hadoop Map Reduce: A Survey. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16). ACM, New York, NY, USA, , Article 143 , 6 pages. DOI: http://dx.doi.org/10.1145/2905055.2905203

[32] Foto Afrati, Shlomi Dolev, Ephraim Korach, Shantanu Sharma, and Jeffrey D. Ullman. 2016. Assignment Problems of Different-Sized Inputs in MapReduce. *ACM Trans. Knowl. Discov. Data* 11, 2, Article 18 (December 2016), 35 pages. DOI: https://doi.org/10.1145/2987376

[33] Java Home Page, http://www.oracle.com/technetwork/java/javase/downloads/index-jsp-138363.html

[34] The Eclipse Foundation open source community website, https://eclipse.org/

[35] Tibco Spotfire home page, http://spotfire.tibco.com/