IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

A Survey on E-Commerce Log Analysis Using Hadoop

Sapna Bhavsar[1*], Pooja Shah[2], Tushar Trambadiya[3]

[1,2]Computer Engineering Department, Shankar Sinh Vaghela Bapu Institute of Technology, Gujarat Technological University, Gandhinagar, India
[3]Information Technology Department, Shankar Sinh Vaghela Bapu Institute of Technology, Gujarat Technological University, Gandhinagar, India

Corresponding author: bhavsarapna1521@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.486489 | Available online at: www.ijcseonline.org

Accepted: 05/Mar/2019, Published: 31/Mar/2019

***Abstract*—** today web mining is a testing assignment in association. Each association produced immense measure of information from different source. Log documents are kept up by the web server. The testing undertaking for E-trade organizations is to know their client conduct to enhance the business by breaking down web log records. Internet business site can produce several peta bytes of date in their web log documents. The investigation of log documents is utilized for learning the client conduct in E-trade framework. The examination of such substantial web log documents requires parallel handling and dependable information stockpiling framework. The Hadoop structure gives solid stockpiling by Hadoop Distributed File System and parallel handling framework for huge database utilizing MapReduce programming model. These components help to process log information in parallel way and figures results productively.

***Keywords*—** Hadoop, MapReduce, Web Log, E-trade, frequent item set mining

I. INTRODUCTION

In this advanced time of huge information, the shopping action is adjusted a great deal in view of huge development in web based shopping sites; it's called E-Tailers. The new age clients want to shop through these online entrances due to different attractions in nations like India, for example, simple and shoddy accessibility of the Internet. The essential reason is extreme challenge between telecoms, for example, Reliance Jio prime enrolment offers free boundless web information utilization for three months for the majority of its clients in ostensible charges. A portion of alternate reasons incorporate worthwhile money back and simple returns without finding of transportation charges from entryways like PayTm, Cash on delivery type customary highlights from E-Commerce locales like Flipkart, Amazon, and other E-Tailers.

II. LITRATURE REVIEW

In the advanced period of enormous information, the shoppers are presently managed about their decision in shopping in E-Trade. Presently Days, E-Trades are utilized diverse Designated ways to deal with snatch the business rewards and For Better Customer Experience.

A. Review Of Recommendation Algorithms based on search methods in E-Trades

The correlation of various Recommendation algorithms and gives a rule to pick the best algorithm. Better and fitting proposals can be achieved by breaking down a lot of information by using Hadoop Map Reduce structure, alongside suitable GUI. All most all the substantial E-Commerce organizations, for example, Amazon, Netflix, and eBay have different types of customized Recommendation framework in changing units. It is utilized for channel and sort item dependent on client's shopping propensities and standards of conduct. The Hybrid Recommendation system will improves the Efficiency and Quality in recommendation. Recommendation Algorithm Module of hybrid recommender system includes content-based recommendation algorithm and collaborative filtering-based algorithm .Personalized recommendation model mainly includes three modules: user modelling, product modelling, and recommendation algorithm.

B. Review Of search methods based on Revamped Market Basket Analysis

The approach is focused on implementing Revamped Market Basket Analysis based on FP-Growth using Hadoop Apache Spark framework which provides rich set of machine learning libraries for machine learning algorithms and studies the performance of the algorithm in spark framework using

Different datasets like retail, mushroom, food art. The performance of the spark implementation is high in terms of execution speed and also with growing dataset size which is compared with BigFIM algorithm implemented in Hadoop. This clearly shows that in order to deal with large datasets Hadoop MapReduce approach is used, but MapReduce has Latency problem which can be overcome using Apache Spark.

C. Review Of Hadoop MapReduce based search method and analysis in E-Trade

A detailed view of processing big data such as Recipe log file with one tera bytes of logs using Hadoop frame work. This paper shows how to process log file using MapReduce and how Hadoop framework is used for parallel computation of log files. Data collected from various resources are loaded into HDFS for facilitating MapReduce and Hadoop framework. We proved that processing big data with the help of Hadoop environment leads to minimum computation and response time and also our HM_PP algorithm leads to good accuracy in prediction of user preferred pages. So, one can easily access the ecommerce system with the help of big data analytics tools with less response time and good prediction accuracy. In future log analysis can be done by correlation engines like RSA envision and HA cloud environment. The above work can also be extended with semantic analysis for better prediction.

D. Review Of Personalised Search Based on Meta search using Relevancy Vector Algorithm

The Approach is to discover uncover that ordinary pursuit frameworks have not advanced to help huge information examination as required by current E-Commerce condition. This work means to create and actualize second-generation HDFS-MapReduce based inventive page Ranking Algorithm, for example Significance Vector (RV) calculation. This examination furnishes the client with a vigorous Meta search instrument, for example IMSS-AE tool to effortlessly comprehend customized look prerequisites and buy inclinations of client.

III. THEORICAL BACKGROUND

A. Hadoop MapReduce

HADOOP [5] is an Apache Software Foundation plot that fundamentally gives two parts:

• A dispersed document framework called HDFS (Hadoop Distributed File System)

• A system and API for building and running MapReduce employments.

HDFS is sorted out like standard UNIX document framework except for that information accumulating is scattered over hardly any machines.

A Map Reduce work generally, parts the informational collection into free units which are dealt with by the guide errands in a thoroughly parallel manner. The structure sorts the yields of the maps, which are then added to the diminish errands. Commonly both the information and the yield of the undertaking are placed in a record framework.

B. Frequent Item set Mining

Frequent set mining [6] assumes a noticeable job as an Information mining errand. It is a technique to discover item sets from the exchange database which happen much of the time. The item set is said to be visit in the event that it happens in least number of transactions. Finding regular thing sets help retail businesses in making vital plans about future trends. Strong association rules are required to uncover the Components that happen together in the informational indexes.

IV. METHODOLOGY

A. Method based on Ranking Comparison and Proposed Deployment Platforms

A positioning correlation of different conceivable huge information organization systems on various qualities, for example, scaling, fault tolerance. Here Rank - 1 demonstrates the best alternative and Rank-5 for most exceedingly awful choice among the majority of the recorded platforms. In general, enormous information applications, there is an exchange off among Scaling and Real-Time Processing capacities [1].

B. Method based on Using Hadoop MapReduce

A new approach for pre-processing of web log and the association rules are being used for extracting patterns from web. In, the first phase discovers user interest patterns and pre-processing is done. Web log analyzer tool [4] is used for analyzing usage pattern from web log. They demonstrated those devices are useful to web head so as to enhance the site execution through the upgrades of substance, structure, introduction and conveyance.

C. Method based on Market Basket Analysis

A new approach for pre-processing of web log and the association rules are being used for extracting patterns from web. In, Market-Basket Analysis [3] is a procedure to distinguish the things that are purchased together because of which; a retail location administrator can settle on vital choices on promoting plans. Most basic significance of the market container investigation is in recommender frameworks which results in expanded benefits in e-commerce. Market-Basket Analysis dependent on FP-Growth is mostly for quicker regular thing set age which can be viewed as a progressive methodology in huge information field.

V. FINDING AND ANALYSIS

An analysis is found based on the method using Hadoop MapReduce. In that, To figure the absolute number of Recipe dependent on cook time gotten by every Recipe Item, a solitary hub Hadoop group is set up with the setups of Ubuntu 14.04 working framework, Hadoop variant 2.6.0, and single hub bunch 192.168.2.1 and dataset Amazon Recipe Logs of 1 Terabyte. Analysis of Recipe thing dependent on cook is appeared in figure 1.

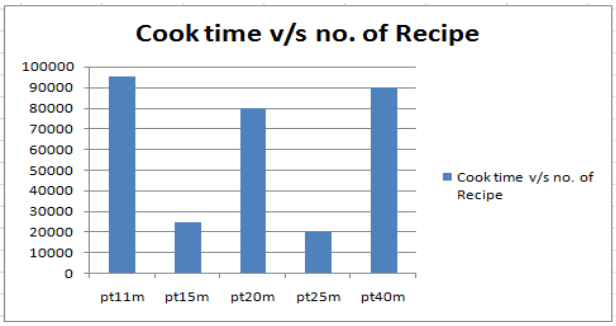

Figure. 1 Analysis of Recipe item based on cook time [2]

An analysis is found based on the method using Market Basket Analysis.In that, the examination was done in a solitary hub Hadoop bunch with Hadoop version 2.6.0. Start was set up on Hadoop with form 1.6. Start gives a rich machine learning library, MLLib, which has different worked.

In machine learning capacities. It has worked in capacities for regular item set mining utilizing FP-Growth. This makes it simpler to build up a quicker visit item set mining calculation in speedier condition. Figure 2 demonstrates a line chart indicating execution time of FPGrowth in start for three datasets: mushroom, retail and foodmart. And the results are, least help is set to half, 40% and 20% for mushroom, retail and foodmart dataset individually. Normal execution time of FP-Growth in Spark is 8 seconds.

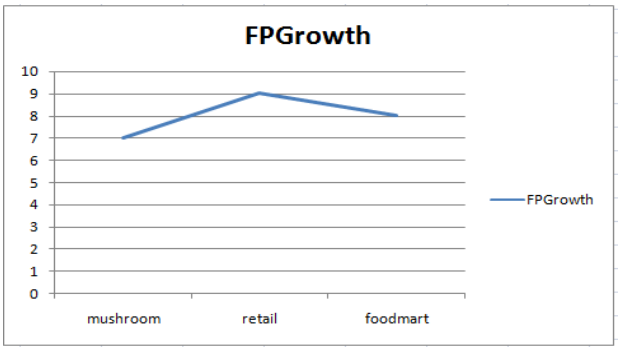

Figure. 2. Performance of FP-Growth in Spark on Different Datasets [3]

An analysis is founded based on the method based on ranking algorithm and based on proposed deployment models Analysis.In that, The customized pertinence of an E-Commerce site to an explicit client for a given item inquiry relies on its position in the yield of query items. To analyze the IMSS-SE Tool with other well-known pursuit apparatuses, Precision of hunt at X metric is considered, which is here appeared by P (X). Different hunt devices utilized for correlation in this Analysis are Meta search tool, web index furthermore, look registry. To assess the productivity and adequacy of proposed RV calculation and IMSS-AE tool. Figure. 3 shows the enhancement in different accuracy parameters at a lot quicker pace when a customized hunt is practiced with proposed IMSS-AE over other expert and well known web indexes, i.e., Google, Yahoo and Meta web indexes, i.e., Dogpile, IMSS-SE.

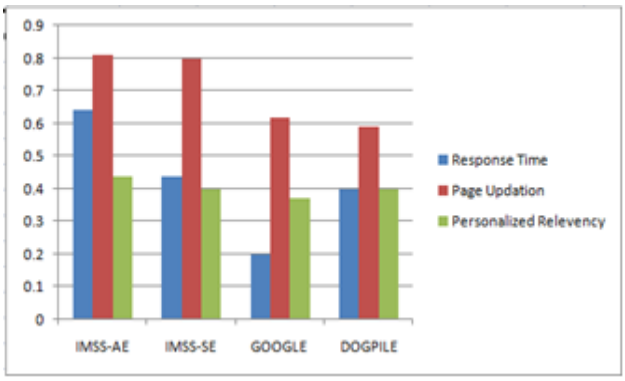

Figure. 3. Search Precision Comparison between IMSS-AE with Dogpile, IMSS- SE, and Google. [1]

VI. CONCLUSION

Hadoop MapReduce Based Personalised E-Commerce search Framework for the second generation Big Data Analytics. IMSS-AE tool will demonstrate the effectiveness of proposed approach over conventional and professional Page Ranking Methods. The execution of the start usage is high as far as execution speed and furthermore with developing dataset measure which is contrasted and BigFIM calculation actualized in Hadoop. MapReduce has inertness issue which can be beaten utilizing Apache Spark. Hadoop structure is utilized for parallel calculation of log records. Information gathered from different assets are stacked into HDFS for encouraging MapReduce and Hadoop system.

ACKNOWLEDGMENT

The Authors gratefully perceives the duties Sapna Bhavsar, Prof. Pooja Shah and Prof. Tushar Trambadiya for their work on the principal type of this record.

REFERENCES

[1] Malhotra, Dheeraj, and O. P. Rishi. "An intelligent approach to design of E-Commerce metasearch and ranking system using next-

generation big data analytics." Journal of King Saud University-Computer and Information Sciences (2018).

[2] Chavan, Reshma, and Debajyoti Mukhopadhyay. "A comparative study of recommendation algorithms in e-commerce." I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017 International Conference on. IEEE, 2017.

[3] Prasad, HR Manjunath. "Revamped Market-Basket Analysis using In-Memory Computation framework." Intelligent Systems and Control (ISCO), 2017 11th International Conference on. IEEE, 2017.

[4] Suguna, S., M. Vithya, and JI Christy Eunaicy. "Big data analysis in e-commerce system using HadoopMapReduce." Inventive Computation Technologies (ICICT), International Conference on. Vol. 2. IEEE, 2016.

[5] Apache Hadoop : ht!p:/lhadoop.apache.org

[6] R.Agrawal and R.Shrikant.Fast Algorithms for mining association rules in large database. In Proc. VLDB, pages 487-499, 1994.

[7] Malhotra, D., Malhotra, M., Rishi, O.P., 2017. An Innovative Approach of Web Page Ranking Using Hadoop- and Map Reduce-Based Cloud Framework. In: Proceedings of Advances in Intelligent Systems and Computing, Vol. 654, CSI, Springer, pp. 421–427.

[8] Malhotra, D., Rishi, O.P., 2017. IMSS: A Novel Approach to Design of Adaptive Search System Using Second Generation Big data Analytics. In: Proceedings of International Conference on Communication and Networks, Springer, pp. 189–196.

[9] Verma, N., Singh, J., 2017. An intelligent approach to Big Data analytics for sustainable retail environment using Apriori-MapReduce framework. Ind. Manage. Data Syst. 117(7), Emerald, 1503–1520..

[10] Verma, N., Singh, J., 2017. A comprehensive review from sequential association computing to Hadoop MapReduce parallel computing in a retail scenario. J. Manage. Analytics, Taylor and Francis. doi:10.1080/23270012.2017.1373261

[11] Wang, H., Wong, K., 2014. Personalized search: An interactive and iterative approach. In Services (SERVICES), 2014 IEEE World Congres, IEEE, pp. 3–10.

[12] Gole, Sheela, and Bharat Tidke. "Frequent itemset mining for Big Data in social media using ClustBigFIM algorithm", 2015 International Conference on Pervasive Computing (ICPC), 2015.

[13] Jiawei Han. 2005. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[14] M.Santhanakumar and C.Christopher Columbus, "Web Usage Analysis of Web pages UsingRapidminer", WSEAS Transactions on computers, EISSN: 2224-2872, vol.3, May 2015.

Authors Profile

Miss. Sapna Bhavsar is Completed her bachelor of Engineering from Gujarat Technological University in 2017.she is pershuing Master Of Engineering in Shanker sinh Vaghela Bapu Institute Of Technology.Her reserch area include Big data Analytics.

Mrs. Pooja Shah has been working as an assistant professor in computer engineering department at Shankersinh Vaghela bapu Institute of Technology. She had completed .her bachelor of engineering from Gujarat University, Gujarat in 2005. After she had completed her master of technology in computer science engineering from Jodhpur National University, Rajasthan in 2012. She is having total 13 years teaching experience at a bachelor and master level along with that she had authored 2 National and 3 International Research papers as a primary and secondary author. Her research areas include networking, information security.

Mr. Tusharkumar Trambadiya has been working as an assistant professor in Information Technology Department at Shankersinh Vaghela bapu Institute of Technology. He had completed his Bachelor of Engineering in IT from North Maharashtra University, Maharashtra in 2010. After he had pursued his master of technology in Computer Science Engineering from Rajiv Gandhi Proudyogiki Vishwavidhyalaya, Bhopal, MadhyaPradesh in 2013. He is having total 6 years teaching experience at a bachelor and master level along with that he had authored 4 International Research papers as a primary and secondary author. He is having keen interest in the field of Big data, Stream data mining, networking.

