PREDICTING REVIEW RATINGS FOR PRODUCT MARKETING

D. Vetriselvi[1], D. Monisha[2], M. Monisha varshini[3]

[1]Assistant Professor, Computer Science Department, JEPPIAAR SRR Engineering College, Tamil Nadu
[2,3]UG Student, Computer Science Department, JEPPIAAR SRR Engineering College, Tamil Nadu

---***---

Abstract - *- In today's developed world, every minute, people around the globe express themselves via various platforms on the Web. And in every minute, a mass amount of unstructured data is generated. This data is in the form of text which is gathered from forums, social media websites , reviews. Such data is termed as big data. User opinions are related to a wide range of topics like on particular products also. These reviews can be mined using various technologies and are of at most importance to make predictions since they directly convey the viewpoint of the masses. Online reviews also have become an important source of information for users before making an informed purchase decision. Early reviewers ratings and their received helpfulness scores are likely to influence product popularity.The challenge is to gather all the reviews ,also calculate and analyse the ratings ,in order to find a refined product ,that scores high rating.*

KeyWords: **Hadoop, HDFS, Hive, Pig, Mysql, Mapreduce, Bigdata**

1. INTRODUCTION

To dissect the attributes of early analysts, we take two essential measurements related with their surveys, i.e., their audit evaluations and supportiveness scores appointed by others. We have discovered that an early analyst tends to appoint a higher normal rating score to items; and an early analyst tends to post more accommodating audits. Our above discoveries can discover importance in the great standards of identity factors hypothesis from sociology, which predominantly contemplates how advancement is spread after some time among the members prior adopters have a more ideal state of mind toward changes than later adopters; and prior adopters have a higher level of conclusion initiative than later adopters.

We can relate our discoveries with the identity factors hypothesis as takes after: higher normal rating scores can be considered as the positive state of mind towards the items, and higher support votes of early audits given by others can be seen as an intermediary proportion of the assessment administration. We additionally clarify this finding with the group conduct broadly considered in financial matters also, human science. Crowd conduct alludes to the truth that people are emphatically affected by the choices of others. To anticipate early commentators, we propose a novel methodology by survey audit posting process as a multiplayer rivalry amusement.

Just the most focused clients can progress toward becoming the early analysts' writes to an item. The opposition procedure can be additionally disintegrated into various pair wise correlations between two players. In a two-player rivalry, the victor will beat the failure with a prior timestamp Past examinations have very accentuated the marvel that people are emphatically impacted by the choices of others, which can be clarified by crowd conduct. The impact of early surveys on ensuing buy can be comprehended as an uncommon case of grouping impact. Early audits contain imperative item assessments from past adopters, which are significant reference assets for ensuing buy choices. As appeared in, when shoppers utilize the item assessments of others to appraise item quality on the Internet, crowd conduct happens in the web based shopping process.

Unique in relation to existing examinations on crowd conduct, we center on quantitatively dissecting the general attributes of early analysts utilizing huge scale certifiable datasets. In expansion, we formalize the early analyst forecast undertaking as an opposition issue and propose a novel installing based positioning way to deal with this undertaking. As far as anyone is concerned, the undertaking of early analyst forecast itself has gotten next to no consideration in the writing.The challenge is to gather all such relevant data, detect and summarize the overall high review ratings on a paricular product.

2. RELATED WORK

Ting Bai, Jian-Yun Nie[1] provided a an early reviewer tends to assign a higher average rating score; and (2) an early reviewer tends to post more helpful reviews. Our analysis of product reviews also indicates that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity. In viewing review posting process as a multiplayer competition game, we propose a novel margin-based embedding model for early reviewer prediction.Experimenting on two different e-commerce datasets have shown that our proposed system outperforms a number of competitive baselines.

Julian McAuley, Alex Yang[2] Provided a Online audits are regularly our first port of call while considering items and buys on the web. While assessing a potential buy, we may have a particular inquiry as a main priority. To answer such inquiries we should either swim through colossal volumes of buyer audits planning to discover one that is pertinent, or generally suggest our conversation starter straightforwardly to the network by means of a Q/A framework. In this paper we would like to meld these two ideal models: given a huge volume of beforehand addressed questions about items, we trust to consequently realize whether an audit of an item is significant to a given question. We define this as a machine learning issue utilizing a blend of-specialists compose system—here each audit is a 'specialist' that gets the opportunity to vote on the reaction to a specific question; all the while we take in an importance capacity with the end goal that 'applicable' audits are those that vote accurately. At test time this scholarly importance work enables us to surface audits that are important to new questions on-request.

Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts [3] provided Collaborative filtering has proven to be valuable for recommending items in many different domains. Here, we explore the use of collaborative filtering to recommend research papers, using the citation web between papers to create the ratings matrix. We tested the ability of collaborative filtering to recommend citations that would be suitable for additional references to target a research paper. We investigated six algorithms for selecting citations, evaluating this through offline experiments against a database of over 186,000 research papers contained in Research Index. We also performed an online experiment with over 120 users to gauge user opinion of the effectiveness of the algorithms and of the utility of such recommendations for common research tasks. We came across large differences in the accuracy of the algorithms in the offline experiment, especially when balanced for coverage. In the online experiment, users felt they received quality recommendations, and were enthusiastic about the idea of receiving recommendations in this domain.

Julian McAuley, Christopher Targett, Qinfeng ('Javen') Shi, Anton van den Hengel[4] intrigued here in revealing connections between the appearances of sets of items, and especially in displaying the human idea of which objects supplement each other and which may be viewed as satisfactory options. We accordingly try to demonstrate what is an on a very basic level human idea of the visual connection between a couple of articles, as opposed to just displaying the visual similitude between them. There has been some enthusiasm generally in displaying the visual style of spots, and objects. We, interestingly, are not looking to show the individual appearances of objects, yet rather how the presence of one question may impact the attractive visual characteristics of another.

Daichi Imamori , Keishi Tajima [5] provided approach for concept Due to the dynamicity, new well known records consistently show up and vanish in miniaturized scale blogging administrations. Early identification of new records that will wind up mainstream in future is an essential issue that has a few applications, for example, slant location, viral showcasing, and client suggestion. Estimation of prominence of a record is additionally valuable for approximating the nature of data it posts. Estimation of the nature of data is vital in numerous applications, yet it is for the most part hard to gauge it without human mediation. Comparative thought has additionally been effectively connected to small scale web journals with connecting capacities. These certainties demonstrated that there is high relationship between the notoriety and the nature of data. In this manner, the estimation of forthcoming notoriety of new records, which have not yet settled the prevalence they merit, is additionally helpful for estimation of the quality.

3. EXISTING SYSTEM

The system used for review extraction and analysis is MySQL S which is a relational database management system. RDBMS uses relations or tables to store E-Commerce data as a matrix of rows by columns with primary key. With MySQL language, E-Commerce data in tables can be collected, stored, processed, retrieved, extracted and manipulated mostly for business purpose. Existing concept deals with providing backend by using MySQL which contains lot of drawbacks i.e data limitation is that processing time is high when the data is huge and once data is lost we cannot recover so thus we proposing concept by using Hadoop tool.

The other drawbacks of the existing system are as follows

- It takes lot of time for performing analysis in large amount of data
- It may result in system failure when a large amount of data is passed to the system
- Major functions and operations for analysis like stemming and NLP processing takes a huge amount of time.

In order to overcome these drawbacks we have proposed the following system with hadoop integration.

4. PROPOSED SYSTEM

The new system is expected to give better performance than the existing system. In our system, an ecommerce mode has huge amount of data related to mode of mobiles, number of features based and range of price vary by finding historical data. The proposed model intension is to develop a model for the mobile data to provide platform for new analytics based on the following queries. The problem they faced till now, they have ability to analyze limited data from databases. In this paper ,we have used both approaches for comparing the results of both the system the first 3 columns of our graph are non hadoop tabs i.e. operations performed on these columns don't use hadoop. Whereas the same operations are performed using hadoop, mapreduce component from terminal, to check the difference in the execution time. For visualizing the output of sentiment analysis in form of pie chart for the particular product which also displays the total review ratings for which the ratings are calculated, a graph chart for comparing the review ratings and the last tab shows the output file from which the review ratings are calculated.

5. SYSTEM ARCHITECTURE

The following diagram shows the overall architecture if the proposed system in details:

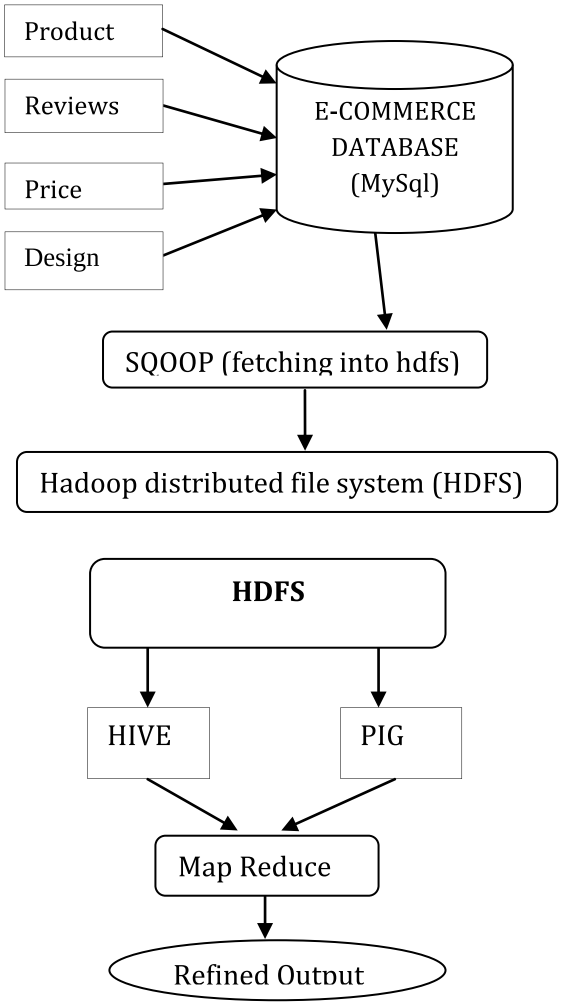
Product
Reviews
Price
Design
E-COMMERCE DATABASE (MySql)
SQOOP (fetching into hdfs)
Hadoop distributed file system (HDFS)
HDFS
HIVE
PIG
Map Reduce
Refined Output

Fig -1: System architecture

The overall system is divided into 5 modules for better and efficient development. Firstly all the reviews are fetched from a blog and stored in a .csv file which is used as an input for the system, the reviews are fetched and stored in a particular format.

In the initial phase, gathering of all review data from different sources as unstructured data. And then, the unstructured data are converted to structured data by using specified techniques like TREC. Now, use those structured data into hadoop framework. Also, analyze data for the following queries:

- ✓ List of different mobiles used in product wise.
- ✓ Prediction of recent trending mobiles are being used periodically with academic year
- ✓ And how many of them are above the range by mention specifically.
- ✓ From this data, which mobiles will be suitable as well as worthable to list out categorical.

The output is shown in a bar chart for a single product and as a bar chart for comparing reviews of multiple product, this is done with the help of the output of hadoop. The initial phase is also done using hadoop to fetch data in HDFS ;In hive andpig,partitions , bucketing and ordering are done.In map reduce component wherein the operations of mapping are done and removed in the reducer phase. The hadoop framework also allows us to increase the efficiency of the system and helps use reduce the overall system time.

TABLE -1: Time taken with and without hadoop

File size	With hadoop	Without Hadoop
>5mb	25s	45s
50mb-70mb	55s	75s
100mb	113s	460s
1gb	1198s	1957s

6. EXPERIMENTAL RESULT AND ANALYSIS

The system is expected to give accurate result for analysis of the sentiments in the form of pie charts and graphs the system uses two approaches to solve the problem using the normal approach and another using hadoop component. The one with hadoop component is expected to be more efficient and faster as compared to the normal system in comparison with large amount of data. This system have given 3 types of input file size small medium and large. The same inputs are passed and processed using normal approach and hadoop integrated approach. The following graph shows the comparison of both the outputs.

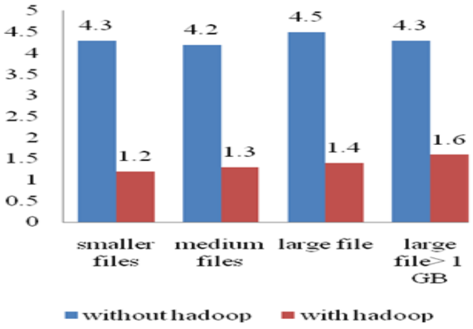

Fig -2: Time taken for with and without hadoop

The above graph we can conclude that for smaller files ,medium files ,large files and greater than 1 gb files in the normal system is quite capable and produces far better results with the hadoop based system as map reduce takes a lot of time for computations as it needs to initialize mappers and reducers andthen combine them into one again for generating final result.

The above graph and the following table describes and summarizes the results from above analysis, Since the data generated and fetched from twitter is in gbs or tbs hadoop will surely give the upper hand in execution from the local machines. The last column in the bar chart describes a .txt file that is more than 1gb in size is really a huge file it may contain cores of tweets and the normal system will crash 9 out of 10 times during executing these dataset hence hadoop ecosystem is used. The hadoop ecosystem will not only prevent the system from crashing but it will also produce results faster and more efficiently.

The above table summarizes the execution time of various file size with and without hadoop integration. As twitter is the largest source of data generation in and around the globe hadoop integration will certainly benefit the analysis and provide is efficient and faster results.

7. SCOPE OF THE SYSTEM

In this paper, we are analyzing E-Commerce data by using hadoop tool along with some hadoop ecosystems like hdfs, mapreduce, sqoop, hive and pig. By using these tools processing of data without any limitation is possible, no data lost problem, we can get high throughput, maintenance cost also very less and it is an open source software, it is compatible on all the platforms since it is Java based.

8. FUTURE WORK

The field of computer science is an endless cycle of development and with every upgraded version there comes a better and more efficient technology. As this system uses hadoop framework, hadoop being a open source development framework sees the ever increasing development with it, in this system we have used hadoop map reduce framework for performing analysis and the hadoop jar has to be processed separately and output needs to be passed to the system, this same can be avoided by using SPARK one of the latest component of hadoop framework.

Apache Spark is an open source processing engine built around speed, case of use, and analytics. If you have large amounts of data that requires low latency processing where a typical Map Reduce program cannot provide, Spark is the alternative. Spark provides in-memory cluster computing for lightning them fast, speed, and supports Java, Scala, and Python APIs for ease of development.

Using SPARK will enable us to reduce the execution time by 10x-100x as SPARK uses the concepts of RDD and the manual execution of the JAR file for map reduce framework would also be avoided. This will make the system more user friendly automated and faster and also more efficient.

9. CONCLUSION

In this paper, we presented a study on E-commerce data and prediction regarding research paper about mobile product. To analysis the E-Commerce data in hadoop ecosystem to improve the business based on number of product sold. Hadoop ecosystem is having hive, pig, mapreduce tools for processing whether output will take less time to process and result will be very fast. Hence in this project already E-Commerce data which is traditionally going to store in RDBMS due to less performance hence by using hadoop tool faster and efficiently processing the data.

REFERENCES

1) Ting Bai, Jian-Yun Nie, "Characterizing and Predicting early reviewers for effective product marketing on e-commerce websites" Journal of Marketing, vol. xx, no. 2, pp. 31 – 38, 2018.

2) J. McAuley and A. Yang, "Addressing complex and subjective product-related queries with customer reviews," in WWW, 2016, pp. 625–635.

3) W.D. J. Salganik M J, Dodds P S, "Experimental study of inequality and unpredictability in an artificial cultural market," in ASONAM, 2016, pp. 529–532.

4) J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Imagebased recommendations on styles and substitutes," in SIGIR, pp. 43–52,2015.

5) D. Imamori and K. Tajima, "Predicting popularity of twitter accounts through the discovery of link-propagating early adopters," in CoRR, 2015, p. 1512.

6) R. Peres, E. Muller, and V. Mahajan, "Innovation diffusion and new product growth models: A critical review and research directions," International Journal of Research in Marketing, vol. 27, no. 2, pp. 91 – 106, 2010.

7) L. A. Fourt and J. W. Woodlock, "Early prediction of market success for new grocery products." Journal of Marketing, vol. 25, no. 2, pp. 31 – 38, 1960.

8) B. W. O, "Reference group influence on product and brand purchase decisions," Journal of Consumer Research, vol. 9, pp. 183–194, 1982.

9) E. M.Rogers, Diffusion of Innovations. New York: The Rise of High- Technology Culture, 1983.

10) K. Sarkar and H. Sundaram, "How do we find early adopters who will guide a resource constrained network towards a desired distribution of behaviors?" in CoRR, 2013, p. 1303.