IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

An Decision Support System to Analyse Customer Behaviour Based on Map Reduce Based C4.5 Algorithm

Mausami Verma[1], Neha Chandrakar[2] , Rahul Kumar Chawda[3]

MCA Student, CS Department, Kalinga University, New Raipur, India[1,2]
HOD, CS Department, Kalinga University, New Raipur, India [3]

Abstract: Big data is one of the most raising technology trends that have the capability for significantly changing the way business organizations use customer behaviour to analyse and transform it into valuable insights. Also decision trees can be used efficiently in the decision making analysis under uncertainty which provides a variety of essential results. Customer behaviour analytics have implemented in many systems, though still it's a developing and unexplored market has greater potential for better advancements. One of the major challenges for knowledge discovery and data mining systems stands in developing their data analysis capability to discover out of the ordinary models in the data. The current work proposes an implementation of C4.5 algorithm using Map Reduce. The experimental results reveal that C4.5 algorithm performs better in case of prediction and accuracy for analysing customer behaviour.

Keywords: Big data, Hadoop, Map Reduce, Web mining, C4.5 Algorithm.

I. INTRODUCTION

Data mining is the process of extracting information and knowledge with potential value from lots of data warehouses. This process can include analytic descriptive and predictive mining, former of which mainly analyses and describes general property attribute of data in database while the latter of which indicates analysis and inference based on the former and carries out prediction additionally. As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today, the new data stored in digital media devices have already been more than 92 % in 2002, while the size of these new data was also more than five exabytes. In fact, the problems of analysing the large scale data were not suddenly occurred but have been there for several years because the creation of data is usually much easier than finding useful things from the data. Even though computer systems today are much faster than those in the 1930s, the large scale data is a strain to analyse by the computers we have today. Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can be divided into two major parts: Web Contents Mining and Web Usage Mining. Web Contents Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined that are web content mining, web structure mining and web usage mining.

- Web Content Mining Web content mining is the process of extracting useful information from the contents of web documents and pages. Content data is the collection of facts a web page is designed to contain.

- Web Structure Mining: The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. The analysed web resources contain the actual web site, the hyperlinks connecting these sites and the path that online users take on the web to reach a particular site .

- Web Usage Mining Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications [6].

Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

- Black Box Data: It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- Social Media Data: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- Stock Exchange Data: The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- Power Grid Data: The power grid data holds information consumed by a particular node with respect to a base station.
- Transport Data: Transport data includes model, capacity, distance and availability of a vehicle.
- Search Engine Data: Search engines retrieve lots of data from different databases.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- Structured data: Relational data.
- Semi Structured data: XML data.
- Unstructured data: Word, PDF, Text, Media Logs.

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

The aim of the study is to create or produce various types of graphs which would showcase certain product analysis which would help the retailer with information to understand the purchase behaviour of a buyer. This information will help the retailer to understand the buyer's needs and reorganize the store's layout accordingly, or even attract new buyers.

II. LITREATURE REV IEW

Paper [8] describes the web usage mining and algorithms used for providing personalization on the web. In this paper focused the data pre-processing and pattern analysis on the web and using the association rule mining algorithms. [9] Describe a web mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being tremendously used in web mining. The results obtained prove the robustness of the algorithm proposed in this paper. [10] this paper focused on Web Usage Mining is the user navigation patterns and their use of web resources. The different stages involved in this mining process and with the comparative analysis between the pattern discovery algorithms Apriori and FP-growth algorithm. [11] Implemented the pre-processing techniques to convert the log file into user sessions which are suitable for mining and reduce the size of the session file by filtering the least requested pages using the pre-processing technique. Data Pre-processing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. This paper presents various methods for handling the problems of big data analysis through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis using HDFS [3]. This paper presents a review of various algorithms from 1994-2013 necessary for handling big data set. It gives an overview of architecture and algorithms used in large data sets. These algorithms define various structures and methods implemented to handle Big Data and this paper lists various tools that were developed for analyzing them. It also describes about the various security issues, application and trends followed by a large data set [4]. The paper presents a broad overview of the topic big data mining, its current status, controversy, and forecast to the future. This paper also covers various interesting and state-of-the-art topics on Big Data mining [5].

III.PROBLEM IDENTIFICATION

Some of the major problems of big data analysis is as mentioned below:

- Incompleteness: It refers to missing of certain data for field values of some samples which creates uncertainties during analysis. This must be managed using certain procedures.
- Heterogeneity : Complicated patterns of data creates problems during analysis which must be managed
- Scale and Complexity: Managing large and complex volumes of data is major issue.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

- Timeliness: As the size of data increases it will take more time to analyse. It is a major issue in Big Data analysis.

IV. METHODOLOGY

Apache Hadoop is a software solution for distributed computing of large datasets. Hadoop provides a distributed filesystem (HDFS) and a Map Reduce implementation. A special computer acts as the "name node". This computer saves the information about the available clients and the files. The Hadoop clients (computers) are called nodes. The "name node" is currently a single point of failure. The Hadoop project is working on solutions for this. The Hadoop file system (HDSF) is a distributed file system. It uses an existing file system of the operating system but extends this with redundancy and distribution. HSDF hides the complexity of distributed storage and redundancy from the programmer. Apache Hadoop consists of two sub-projects :

- Hadoop MapReduce : MapReduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

- HDFS (Hadoop Distributed File System): HDFS takes care of storage part of Hadoop applications. MapReduce applications consume data from HDFS. HDFS creates multiple replicas of data blocks and distributes them on compute nodes in cluster. This distribution enables reliable and extremely rapid computations.

- Some of the features of Hadoop are as follows:

1. Suitable for Big Data Analysis:As Big Data tends to be distributed and unstructured in nature, HADOOP clusters are best suited for analysis of Big Data.

2. Scalability: HADOOP clusters can easily be scaled to any extent by adding additional cluster nodes, and thus allows for growth of Big Data.

3. Fault Tolerance: HADOOP ecosystem has a provision to replicate the input data on to other cluster nodes. That way, in the event of a cluster node failure, data processing can still proceed by using data stored on another cluster node.

The flow of the system is as follows:

- Customer dataset will be loaded from the HDFS as input for the algorithm.
- Invoke the instance of C4.5 class.
- Using the MapReduce framework of Hadoop.
- Reduce function counts number of occurrences of combination of and prints count against it.
- Calculate entropy, information gain and gain ratio of attributes. Process the input dataset from HDFS according to the defined algorithm of C4.5 in MapReduce framework.
- Generate the rules and store it in HDFS.
- Provide Recommendation according to their interest.

C4.5 is an algorithm developed by Ross Quinlan that generates Decision Trees (DT), which can be used for classification problems. It improves (extends) the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. Its commercial successor is C5.0/See5, a lot faster than C4.5, more memory efficient and used for building smaller decision trees.

V. EXPERIMENTAL RESULT

By using the C4.5 algorithm the predicted data will help to get higher accuracy than the previous. In this experiment by using lower rate of prediction we can get most accurate values that are by using the lower prediction rate we can recommend items of the customers interest by using the higher accuracy.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

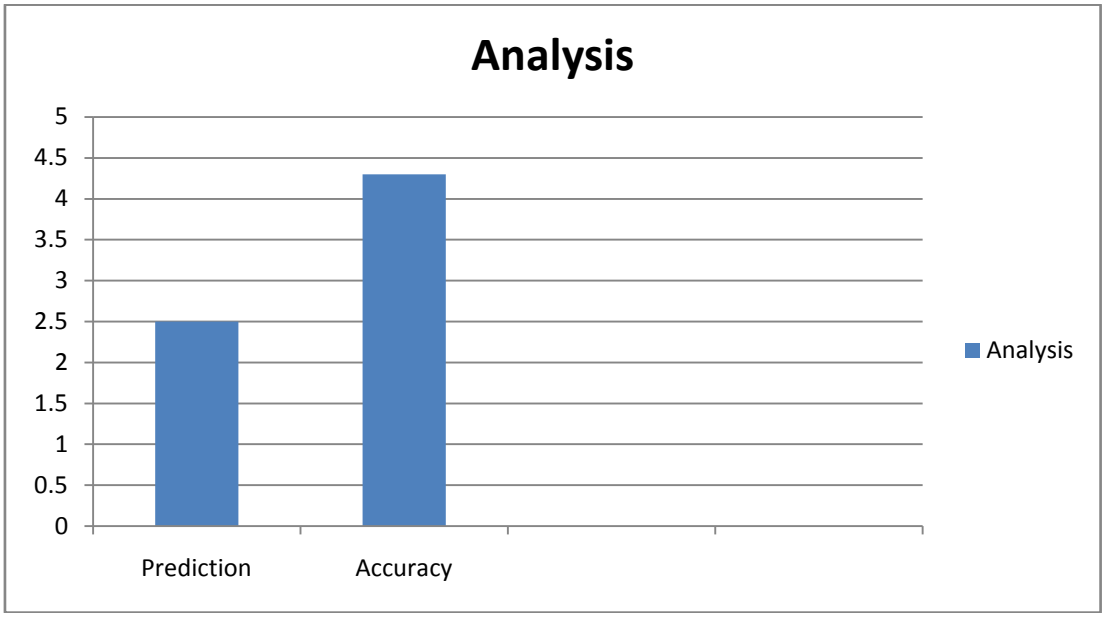

Figure: C4.5 Algorithm Performance Analysis

.

VI. CONCLUSION

Big-Data approaches can be utilized in a very efficient manner in order to analyses the online behavior of customers. The current work conducts a preliminary investigation for developing an efficient approach for developing a decision support system for helping the managers in decision making using Big-Data Map Reduce framework. Once the customer dataset is loaded in the framework the C4.5 algorithm generates a set of rules identified by Decision Trees. This proposed approach is highly efficient for predicting the future of product as per customer behavior. The future work aims to conduct an experiment on utilizing parallel execution of decision support rules for increasing the efficiency.

REFERENCES

[1] Q.Li, J.Xing, O.Liu, and W.Chong "The Impact of Big Data Analytics on Customers" Online Behaviour", Proceedings of the International MultiConference of Engineers and Computer Scientists 2017.
[2] J.K.U and J.M.David,"Issues,Challenges and Solutions:Big Data Mining",NetCoM ,CSIT,2014.
[3] P.Duggal S.Paul."Big Data Analysis : Challenges and Solutions",International Conference on Cloud,Big Data and Trust 2013,Nov 13-15 ,RGPV
[4] C.Yadav,S.Wang and M.Kumar,"Algorithm and Approaches to Handle Large Data – Survey",IJCSN,2013.
[5] W.Fan, A.Bifet, "Mining Big Data: Current Status and Forecast to the Future",2013,SIGKDD
[6] R.Gupta, "Journey from Data Mining to Web Mining to Big Data"
[7] P.B. Mohata and S.Dhande ," Web Data Mining Techniques and Implementation for Handling Big Data", IJCSMC, Vol. 4, Issue. 4, April 2015.
[8] S. Jagan, and S.P. Rajagopalan, "A survey on web personalization of web usage mining", IRJET International Research Journal of Engineering and Technology, 2015.
[9] A. Ladekar, P. Pawar, D. Raikar and J. Chaudhari, "Web Log Based Analysis of User's Browsing Behavior", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.
[10] S. Parvatikar and B. Joshi, "Analysis of User Behavior through Web Usage Mining", ICAST – International Conference on Advances in Science and Technology, 2014. [11] A. Deepa, and P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.
[11] [N. Anand, "Effective prediction of kid's behavior based on internet use", International Journal of Information and Computation Technology, 2014.
[12] ei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, 14 (2), pp1-5 [9] Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) "Algorithm and Approaches to handle large Data- A Survey", IJCSN, 2(3), ISSN:2277-5420(online), pp2277-5420 [12] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data"
[13] Puneet Singh Duggal, Sanchita Paul, (2013), "Big Data Analysis:Challenges and Solutions", Int. Conf. on Cloud, Big Data and Trust, RGPV