Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 79 (2016) 986 – 992

Procedia Computer Science

7th International Conference on Communication, Computing and Virtualization 2016

Performing Customer Behavior Analysis using Big Data Analytics

Anindita A Khade

Assistant Professor,SIESGST NERUL,India

Abstract

Although there are many systems that have implemented customer behavior analytics, it's still an upcoming and unexplored market that has greater potential for better advancements. Big data is one of the most rising technology trends that have the capability for significantly changing the way business organizations use customer behavior to analyze and transform it into valuable insights. Even decision trees can be used efficiently for analyzing data. At the end of this paper, a proposed Map Reduce implementation of well-known statistical classifier, C4.5 decision tree algorithm has been proposed. Apart from this,the system aims to implement Customer data visualization using Data Driven Documents (d3.js) which allows us to build well customized graphics.

Peer-review under responsibility of the Organizing Committee of ICCCV 2016

Keywords: Big Data analyti;C4.5 algorithm;D3.j;,Data visualization;Hadoop;MapReduce

1. Main text

Here Big data is a collection of unstructured data that has very large volume, comes from variety of sources like web ,business organizations etc. in different formats and comes to us with a great velocity which makes processing complex and tedious using traditional database management tools .It can be termed as a growing torrent. So the major demanding issues in big data processing include storage, search, distribution, transfer, analysis and visualization.

Earlier, the term 'Analytics' indicated the study of existing data to research about potential trends and to analyze the effects of certain decisions or events that can be used for business intelligence to gain various valuable insights. Today's biggest challenge is how to discover all the hidden information through the huge amount of data collected from a varied collection of sources. There comes Big Data Analytics into picture. One of them is the customer behavior analysis which is referred as customer analytics.

Peer-review under responsibility of the Organizing Committee of ICCCV 2016
doi:10.1016/j.procs.2016.03.125

Customer analytics helps to turn big data into big value by allowing the organizations to predict the buyer behavior thereby improving their sales, market optimization, inventory planning, fraud detection and many more applications. A wide range of approaches are available and can be implemented but the one that stands out is the use of decision trees for the purpose of classification that can be efficiently used in consumer analytics.

Various decision tree algorithms have been developed over a period of time with enhancement in performance and ability to handle various types of data. One of the well-known decision tree algorithm is C4.5 is C4.5 [3-4], an extension of basic ID3 decision tree algorithm [5]. Customer analytics is incomplete without visualization of the data. In addition to classification of data using decision trees it is also important to visualize the data so that organizations get a visual aspect of the data in order to understand the variations in customer patterns.

2. Literature Survey

Traditional Analytical Systems For Customer Behavior[7]:

In the late 1970s, there were two approaches for constructing Database Management System's (DBMS's). The first approach was based on the hierarchical data model, typified by (Information Management Systems) from IBM, in response to the enormous information storage requirements generated by the Apollo space program. The second approach was based on the network data model, which attempted to create a database standard and resolve some of the difficulties of the hierarchical model, such as its inability to represent complex relationships DBMSs. However, these two models had some fundamental disadvantages like the complex programs had to be written to answer even simple queries. Also there was minimal data independence .

Many experimental relational DBMS were implemented thereafter, with the first commercial products appearing in the 1970's and early 1980's. Relational DBMS used extensively in the 80's and 90's was limited in meeting the more complex entity and data needs of companies, as their operations and applications became increasingly complex. In response to the increasing complexity of database applications, two new data models had emerged, the Object-Relational Database Management Systems (ORDBMS) and Object-Oriented Database Management Systems (OODBMS), which subscribes to the relational and object data models respectively. The OODBMS and ORDBMS have been combined to represent the third generation of Database Management Systems.

Dawn Of Big Data Analytics:

Data turns to big data when its volume, velocity, or variety go beyond the abilities of the IT operational systems to gather, store, analyze, and process it. Most of the organizations are capable of handling vast amount of unstructured data using varied tools and equipments but with the rapidly growing volume and fast flood of data, they do not have the capability of mining it and derive necessary insights in a well-timed way.

Big Data is emerging from the realms of science projects at companies to help telecommunication giants understand exactly which customers are happy with their service and what processes caused the dissatisfaction, and predict which customers are going to change the service. To obtain this information, billions of loosely-structured bytes of data in different locations needs to be processed until the required data is found out. This type of analysis enables executive management to fix faulty processes or people and may be able to reach out to retain at-risk customers . Big data is becoming one of the most important technology trends that have the potential for dramatically changing the way organizations use customer behaviour to analyze and transform it into valuable insights.[11]

Key concepts of Customer analytics[6] :

The survey on customer analytics revealed the following key concepts:

1) Venn Diagram– Discovering Hidden Relationships

Combine multiple segments to discover connections, relationships or differences. Explore customers that have bought different categories of products and easily identify cross-selling opportunities.

2) Data Profiling– Identify Customer Attributes

Select records from your data tree and generate customer profiles that indicate common features and behaviors. Use customer profiles to inform effective sales and marketing strategy.

3) Forecasting – Time Series Analysis

Forecasting enables you to adapt to changes, trends and seasonal patterns. You can accurately predict monthly sales volume or anticipate to the number of orders expected in any given month.
4) Mapping – Identify Geographical Zones

Mapping uses color-coding to indicate customer behavior as it changes across geographic regions. A map divided into polygons that represent geographic regions shows you where your churners are concentrated or where specific products sell the most.

5) Association Rules – Cause/ Effect – Basket Analysis

This technique detects relationship or affinity patterns across data and generates a set of rules. It automatically selects the rules that are most useful to key business insights: What products do customers purchase simultaneously and when? Which customers are not buying and why? What new cross-selling opportunities exist?

6) Decision Tree – Classify and Predict Behavior

Decision trees are one of the most popular methods for classification in various data mining applications and assist the process of decision making. Classification helps you do things like select the right products to recommend to particular customers and predict potential churn. Most primarily used decision tree algorithms include ID3, C4.5 and CART.

Tools for data visualization
Polymaps: Polymaps is a free JavaScript library and a joint project from SimpleGeo and Stamen. This complex map overlay tool can load data at a range of scales, offering multi-zoom functionality at levels ranging from country all the way down to street view. [12]

Flot: A JavaScript plotting library for jQuery, Flot is a browser-based application compatible with most common browsers — including Internet Explorer, Chrome, Firefox, Safari and Opera. Flot supports a variety of visualization options for data points, interactive charts, stacked charts, panning and zooming, and other capabilities through a variety of plugins for specific functionality. [12]

3) D3.js: A JavaScript library for creating data visualizations with an emphasis on web standardsUsing HTML, SVG and CSS, bring documents to life with a data-driven approach to DOM manipulation — all with the full capabilities of modern browsers and no constraints of proprietary frameworks. [12]
4) SAS Visual Analytics:SAS Visual Analytics is a tool for exploring data sets of all sizes visually for more comprehensive analytics. With an intuitive platform and automatic forecasting tools, SAS Visual Analytics allows even non-technical users to explore the deeper relationships behind data and uncover hidden opportunities. [12]

3. Related Technologies

1.1 Apache Hadoop

Apache Hadoop[13] is an open source software framework [16]. Hadoop consists of two main components: a distributed processing framework named MapReduce and a distributed file system known as the Hadoop distributed file system, or HDFS[2]. One of the most important reason for using this framework in this project is to process a large amount of data and do its analysis which is not possible with other system. The storage is provided by HDFS and the analysis is done by MapReduce. Although Hadoop is best known for MapReduce and its distributed file system, the other subprojects provide complementary services, or build on the core to provide high-level abstractions. [1]

1.2 Hadoop Distributed File System:

The Hadoop Distributed File System (HDFS)[15] is the storage component. In short, HDFS provides a distributed architecture for extremely large scale storage, which can easily be extended by scaling out. When a file is stored in HDFS, the file is divided into evenly sized blocks. The size of block can be customized or the predefined one can be used. In this project, the customer dataset is stored in HDFS. The dataset contains a lot of customer records with respect to purchases. Also, the output file containing decision rules of is written into HDFS.

1.3 Map Reduce Model:

MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster. MapReduce works by breaking the processing into two phases: the map phase and the Reduce phase. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the Map function and the Reduce function. The input to our map phase is the raw data of customers. We choose a text input format that gives us each line in the dataset as a text value. The key is the offset of the beginning of the line from the beginning of the file. The output from the map function is processed by the MapReduce framework before being sent to the reduce function. This processing sorts and groups the key-value pairs by key. [1]

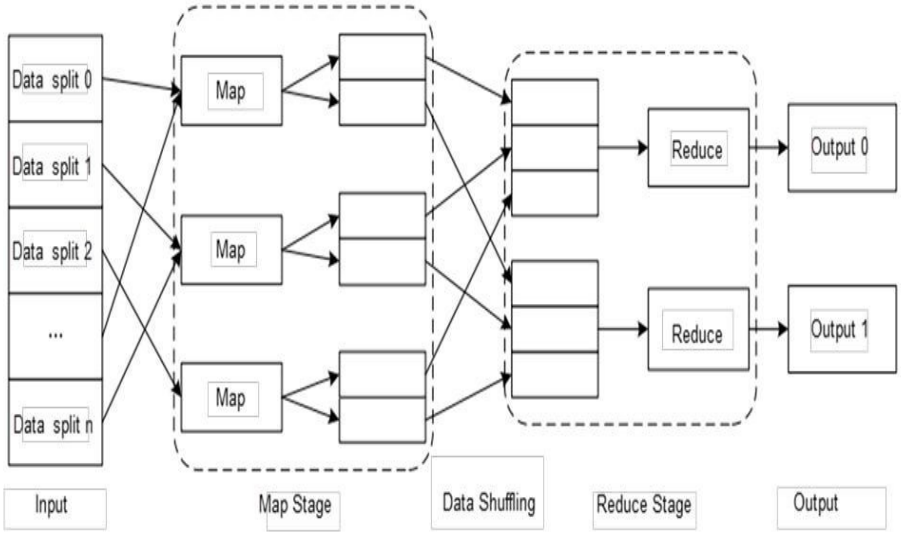

Fig. 1: MapReduce Programming Model

Java code for the map function and the reduce function for this implementation is written

for overriding the default map and reduce function provided by hadoop framework. The programming logic for the respective is based on C4.5 algorithm.

2. *Methodology*

The flow of the system is as follows:

1) Loading the customer dataset from HDFS as input for the algorithm.
2) Invoke the instance of C4.5 class.
3) Using the MapReduce framework of Hadoop, Map function is invoked which checks whether this instance belongs to Current Node or not. For all uncovered attributes it outputs index and its value and class label of instance.
4) Reduce function counts number of occurrences of combination of (index and its value and class Label) and prints count against it.
5) Calculate entropy, information gain and gain ratio of attributes.
6) Process the input dataset from HDFS according to the defined algorithm of C4.5 decision tree data mining in MapReduce framework.
7) Generate the decision rules and store it in HDFS.
8) Accept the new test data from web UI.
9) Access the rules and based on it, decide the category of the new data.
10) Provide visualization of the dataset from HDFS on the Web UI in the form of bar graphs, pie charts etc. using D3.js.

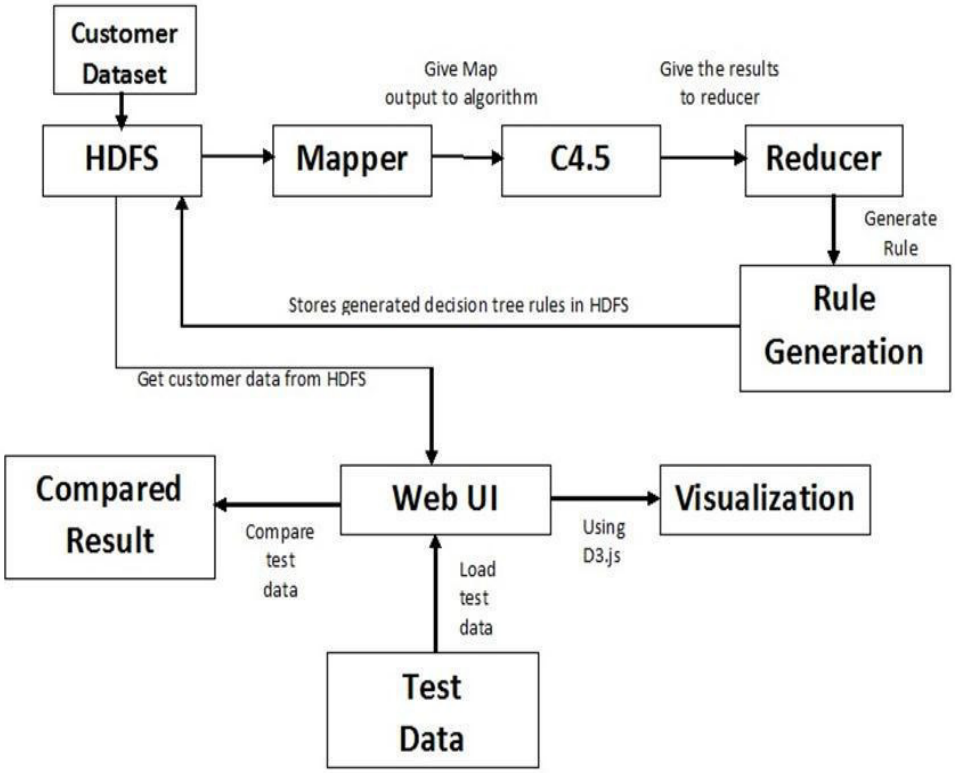

Fig. 2: Flowchart of the Proposed System

2.1 *C4.5 Algorithm:*

C4.5[3-4] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason C4.5 is often referred to as a statistical classifier. C4.5 algorithm uses information gain as splitting criteria. It can accept data with categorical or numerical values. To handle continuous values it generates

threshold and then divides attributes with values above the threshold and values equal to or below the threshold. C4.5algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5.[8]

Let C denote the number of classes. In this case, there are two classes in which the records will be classified into. The classes are yes and no. The p(S, j) is the proportion of instances in S that are assigned to j -th class. Therefore, the entropy of at tribute S is calculated as:

Entropy(S) = -∑ j=1c p(S,j) *log p(S,j)

Entropy is calculated for each record of a particular attribute.
Accordingly the information gain by a training dataset T is defined as:

Gain(S,T)= Entropy(S)-∑v€values(Ts)|T(s,y)/T(s)* log p(S,j) where Values (TS) is the set of values of S in T , Ts is the subset of T induced by S , and TS ,v is the subset of T in which attribute S has a value of v.

2.2 *Data Visualization using D3.js:*

D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and

CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.[14]

Key features of D3.js[14]:

- Bind arbitrary data to DOM
- Create interactive SVG bar charts
- Generate HTML tables from data sets
- Variety of components and plugins to enhance capabilities
- Built-in reusable components for ease of coding

4. Conclusion

This paper defines the proposed system for distributed implementation of C4.5 algorithm using MapReduce framework along with the customer data visualization. With the rise in development of cloud computing and big data, traditional decision tree algorithms cannot fit any more and hence we introduced the mapreduce implementation of C4.5 decision tree algorithm. Visualization done using D3.js is fast and reusable because it uses traditional HTML elements along with Scalable Vector Graphics (SVG). In future works, the use of fast and real time database systems like Apache HBase or MongoDB can be incorporated with this system. In addition to this, we can use distributed refined algorithms like ForestTree implemented in Apache Mahout to increase efficiency and scalability.

5. References

1. Tom white, ―Hadoop - The Definitive Guide‖,3rd.Edition, O'Reilly Media, Inc.,Sebastopol, CA 95472,2012.
2. Dirk deRoos, Paul C. Zikopoulos, Bruce Brown,Rafael Coss, Roman B. Melnyk ―Hadoop For Dummies‖, John Wiley & Sons, Inc., Hoboken, New Jersey,2014
3. J.R. Quinlan,―C4.5: programs for machine learning‖, Morgan Kaufmann,1993.

4. J.R. Quinlan,―Improved use of continuous attributes in C4.5‖, arXiv,1996 ,preprint cs/9603103.
5. J.R. Quinlan,―Induction of decision trees‖,Machine Learning, vol.1, no.1,1986,pp.81-106.
6. Actuate Corporation, ―Customer Analytics Turn Big Data into Big Value‖. Available : http://birtanalytics.actuate.com/customer-analytics-turn-big-data-into-big-value
7. Seamus Rispin, ―Database Resources,‖ The Institute of Certified Public Accountants, Ireland.
8. Mr. Brijain R Patel, Mr. Kushik K Rana (2014).A Survey on Decision Tree Algorithm for Classification. IJEDR [Online]2(1). Available: http://www.ijedr.org/papers/IJEDR1401001.pdf
9. Wei Dai and Wei Ji. (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm. International Journal of Database Theory and Application [Online] 7(1), pp. 49-60. Available: http://www.chinacloud.cn/upload/2014-03/14031920373451.pdf
10. Surbhi Hardikar, Ankur Shrivastava ,Vijay Choudhary(2012)Comparison between ID3 and C4.5 in Contrast to IDS[Online] 2 (7), pp.659-667.Available : www.vsrdjournals.com
11. David Floyer.(2014, Jan) Enterprise Big-data [Online] Available: http://wikibon.org/wiki/v/Enterprise_Big-data
12. Andy Lurie(2014,Feb).39 Data visualization tools for big data[Online]. ProfitBricks,The Laas Company.Available : https://blog.profitbricks.com/39-data-visualization-tools-for-big-data
13. Apache Hadoop http://hadoop.apache.org/releases.html
14. D3.js : http://d3js.org
15. HDFS : http://hortonworks.com/hadoop/hdfs
16. http://en.wikipedia.org/wiki/Apache_Hadoop