Customer Behavior Analysis using Statistical Classifier C4.5 Implemented on Apache Spark over Hadoop YARN

Priya G Nambiar[1], E K Girisan[2]
[1]*M Phil Scholar,* [2]*Associate Professor,*
[12]*Department of Computer Science,*
[1]*Sree Narayana Guru College, Coimbatore,Tamil Nadu, India*
[1]*Bharathiar University*
[2]*Sree Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India*

Abstract- Understanding the customer is the primary goal for any retailer. The most important challenge of online shopping is tostudy and analyze the behavior of the customer. Customer behavior is used to analyze and transfer them into valuable insights. Even though there are many systems that have implemented the customer behavior analytics, it is still an upcoming and unexplored market that has greater potential for better advancement. Decision tree can be used efficiently for analyzing the data. The younger generation focuses more on online shopping and it increases the online sales. This paper provide insights on how the online data can support customized targeting, resulting in increase in ecommerce revenue by advanced predicting modelling on visitor's behavior. This paper proposed a memory based framework like Apache Spark implementation of well-known statistical classifier c4.5 decision tree algorithm. In addition to this, the system aims to implement customer data visualization using Origin Software which allows to build well customized graphics.

Keywords- Big data Analytics, Hadoop, HDFS, Yarn, Spark, RDD, Decision tree C4.5, Origin.

I. INTRODUCTION

Due to the advent of new technologies, devices and social networking sites, the amount of data produced is growing rapidly every year. This rate is still growing enormously. There comes "big data" into picture. BigData really means a big data. BigData is the data sets that are so big and complex that traditional data processing application software are inadequate to deal with. BigData is the new science of understanding and predicting the human behavior by studying and analyzing the large volumes of unstructured data. BigData not only deals with data volume but also with data variety and data velocity. The major demanding issues in big data processing includes storage, search, distribution, transfer, analysis and visualization.

The analytics indicates the study of existing data to research about new trends and patterns and to analyze the effects of certain decision that can be used for business intelligence to gain various valuable insights. Therefore, todays biggest challenge is how to discover all the hidden information from huge amount of data collected through different sources. There comes the picture of Predictive Analytics especially Customer Behavior Analytics.

Predicting the future is not gazing into the crystal ball and seeing what is going to happen next. Customer Analytics are used to predict the behavior of customers, helps to know what kind of products will become successful in particular reason and plan strategy for communicating to customers and thereby improving the sales, market optimization and inventory planning. The key benefit of predicting analytics is to identify more sophisticated and granular insights by analyzing the data set with greater speed.

A wide range of approaches are available and can be implemented but the most popular technique used in organization is the use of decision tree for the purpose of classification that can be efficiently used in Customer analytics.

Various decision tree algorithm has been developed over a period of time but the most well-known decision tree algorithm is C4.5, which has the ability to handle various types and volumes of data. Apache Spark over Hadoop YARN implementation of c4.5 algorithm brings a huge benefit for Customer Analytics. The main feature of spark is its in-memory cluster computing that increases the processing speed of an application.

Apart from this classification of data using decision tree, it is important to visualize the data so that organization get a visual aspect of data in order to understand the variation in customer pattern.

II. RELATED WORKS

Big data is getting larger day by day and data continues to be explode. Considering various valuable customer experiences from large amounts of structured and unstructured data from different sources in different formats, it requires proper structures and tools. To obtain maximum business impact, this process requires proper combination of people, process and analytical tools. In order to establish a long lasting relationship, more tactical ways should be considered for making customer stay, their loyalty and relationships. Recent studies revealed that different approaches have been used by giants for understanding the customer behavior in order to maximize their business. Big data is becoming one of the most important technology trends that have the potential for dramatically changing the way organization use customer behavior to analyze and transform it into valuable insights.

The survey of customer analytics revealed the following key concepts:

1) Data Profiling – Identify the attributes of customer

This approach allows to select records from data tree and generate customer profiles that indicate common features and behaviours and use customer profiles to make effective sales and improvise the marketing strategy.

2) Forecasting – Time Series Analysis

Forecasting enables to adapt to changes, trends and seasonal patterns. This method accurately predicts monthly sales volume or anticipate to the number of orders expected in any given month.

3) Mapping – Identify Geographical Zones

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

4) Association Rules – Basket Analysis

This technique detects affinity pattern across data and generates a set of rules. It automatically selects the rules that are most useful to key business insights: What product does customer buy simultaneously and when? Which customers are not buying and why? What new cross-selling opportunities exist?

5) ough Set Approach

This approach characterizes the behaviour of customer to maintain long-term profitable customers. This concept is used to predict the entrepreneur behaviour The rough set approach focuses on customer segmentation and rule generation for customer behaviour analysis.

6) Hierarchical and Network Approach

This approach effectively analysing customer behaviour using two different model- hierarchical data model and network data model. This model overcome some of the difficulty such as inability to represent complex relationship in the database management system.

7) Click Stream Approach

Click stream approaches extract information and make predictions about customers shopping behaviour and also may provide some hints about their buying behaviour. This model predicts whether customers will or will not buy their items are added to shopping baskets on a digital market place.

8) Decision Tree – Classify and Predict Behaviour

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 to select the right products to recommend to particular customers and predict potential churn. Most primarily used decision tree algorithms are ID3, CART and C4.5.

III. RELATED TECHNOLOGIES FOR PROPOSED MODEL

1) Apache Hadoop

Apache Hadoop is an open source software framework. All the components of Hadoop ecosystems are Hadoop common, Hadoop YARN, Hadoop Distributed File System(HDFS) and Hadoop MapReduce. Hadoop common provides all Java libraries, utilities, OS level abstraction, necessary Java files and script to run Hadoop, while Hadoop YARN is a framework for job scheduling and cluster resource management. HDFS in Hadoop architecture provides high throughput access to application data and Hadoop MapReduce provides YARN based parallel processing of large data sets. The proposed model used Apache Spark over Hadoop YARN instead of Hadoop MapReduce.

2) Hadoop Distributed File System

The default big data storage layer for Apache Hadoop is HDFS. HDFS is the "Secret Sauce" of Apache Hadoop components as users can dump huge datasets into HDFS and the data will sit there nicely until the user wants to leverage it for analysis. HDFS component creates several replicas of the data block to be distributed across different clusters for reliable and quick data access. HDFS comprises of 3 important components-NameNode, DataNode and Secondary NameNode. HDFS operates on a Master-Slave architecture model where the NameNode acts as the master node for keeping a track of the storage cluster and the DataNode acts as a slave node summing up to the various systems within a Hadoop cluster. In this paper, the customer dataset is stored in HDFS. The dataset contains a lot of customer records with respect to purchase. In addition to this the output file containing decision rules is also stored in HDFS.

3) Hadoop YARN

Apache Hadoop YARN is the resource management and job scheduling technology in the open source Hadoop distributed processing framework. One of Apache Hadoop's core components, YARN is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.YARN stands for Yet Another Resource Negotiator. YARN is similar to master slave architecture. Resource Manager is responsible for uniform resource management and schedule. When an application is submitted, an Application Master is needed to track and supervise the job. The architecture of YARN with Spark as the application is shown as below

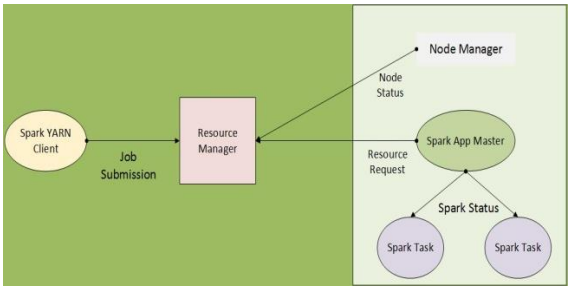

Fig.1: Architecture of YARN with Spark

4) Apache Spark

Apache Spark is an open source distributed computing framework which is designed for low-latency and iterative computation on historical data. Spark provides an easy-to program interface that is available in Java, Python and Scala. Apache Spark over Hadoop YARN is proposed in this paper.

A. Resilient Distributed Databases

Spark provides a fault tolerant and efficient memory abstraction called Resilient Distributed Databases. When a RDD is created, the users can decide which intermediate RDDs are to be kept in memory and control their partitioning to optimize data placement to get high-efficiency result.

B. Operations on RDDs

The operations on RDDs are mainly classified into two categories: transformations and actions. With transformation, the user can create a new dataset from an existing RDD. All transformations in spark are lazy. After the operations of actions, a value is returned to the driver program

C. Job Scheduling

A DAG is built from the RDDs lineage graph when a job is committed to the master of the cluster. DAG consists of several stages. The stages are divided into two categories: shuffle map stage and result stage. Shuffle stages are those that their results are input for another stage while result stage are those that their tasks directly compute the action that initiated a job.

D. Origin Software for Data visualization

Origin Software is typically a computer program used for interactive scientific graphing and for data analysis purpose. It was developed by Origin Lab Corporation and runs on Microsoft Windows. Graphing support in Origin includes various 2D &3D plot types. Data Analysis in Origin include statistics, signal processing, curve fitting and peak analysis. Origin imports data files in various formats such as ASCII text, Excel, SPC etc. It also exports the graph to various image file formats such as JPEG, GIF, TIFF etc. The key features of Origin include it has a scripting language for controlling the software. Origin can be used as a COM server for programs written in visual basic, .NET and c#. Origin project file (.OPJ) can be read by the open source library (liborigin).

IV. OVERVIEW OF DECISION TREE

Decision Tree is one of the most prominent data mining technology. A decision tree is a classification scheme which generates a tree and a set of rules, representing the model of different classes from a given data set. In Decision Tree, every internal node represents test on attribute, every branch represents the output of the test and every leaf node store a class label. Most primarily used decision tree algorithms are ID3, CART and C4.5. The C4.5 algorithms was developed by J. Ross Quinlan in 1993 which was an extension of ID3 algorithm. C4.5 algorithms have been widely used by business giants to maximize the sales and to long-lasting the relationship with their customer.

C4.5 adopts recursive and top-down method to construct a decision tree from the training items and the categories they belong to. The detail procedures are shown as below

1) Initially get the data set of Dset. Each item in Dset has some attribute values and a class label.
2) Then gain ratio is computed from splitting attribute.
3) Create a decision node that splits on att_high where att_high be the attribute with highest gain ratio.
4) After splitting on attt_high, some subcubes are formed. For each cube of ChildCube the subcubes, go back to 2) to get att_high of child. Att_high will be the child of the node formed in step 3).

In addition to this some pruning operations will be performed to overcome the excessive fitting.

The entropy of a dataset to be classified is measured as

$$\boldsymbol{Info(D) = \sum_{i=1}^{m} pi \log_2(pi)} \quad (1)$$

The *pi* refers to the probability of one item that belongs to class *Ci*, and is measured by $|C_{i,D}|/|D|$. *Info(D)* is called the entropy of D. We need to get accurate classification of the dataset except the information *Info(D)* and that is measured as

$$\boldsymbol{Info(D) = \sum_{j}^{v} \frac{|Dj|}{|D|} \times Info(D_j)} \quad (2)$$

The $\frac{|Dj|}{|D|}$ acts as the weight of j[th] partition. $Info_A\ (D)$ is the expected information according to A to classify the items in D. The information gain is defined as the difference between the original information *Info(D)* and the new information $Info_A\ (D)$:

$$\boldsymbol{Gain(A) = Info(D) - Info_A\ (D)} \quad (3)$$

C4.5 uses split information to normalize the information gain:

$$\boldsymbol{SplitInfo_A\ (D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2 |\frac{D_j}{D}|} \quad (4)$$

The standard C4.5 used to split a node is gain ratio, which is shown as follows:

$$\boldsymbol{GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}} \quad (5)$$

V. ARCHITECTURE OF PROPOSED MODEL

In order to predict the behavior of customer on YARN using Spark, some tools and infrastructure are required. The architecture of the proposed model is shown below:

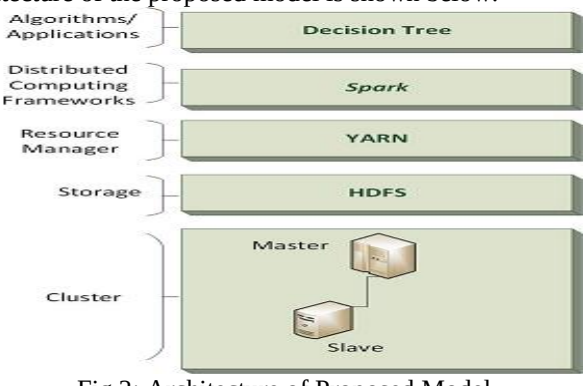

Fig.2: Architecture of Proposed Model

A cluster with some nodes of Ubuntu operating system as the base infrastructure of the whole system. On the top of the infrastructure HDFS (Hadoop Distributed File Systems) is used for data storage. Hadoop provides shell operations and APIs for the users to have access to the data stored on HDFS. On top of HDFS, YARN is chosen as resource manager and applications master to manage the cluster. For data processing and analysis, Spark is selected. With the Spark frames it is possible to develop parallel algorithms or application. Here C4.5 Decision Tree algorithm is chosen.

C4.5 on Spark

As Spark has different APIs and operations it is necessary to write a Spark application. The work flow of Spark is shown in the following diagram

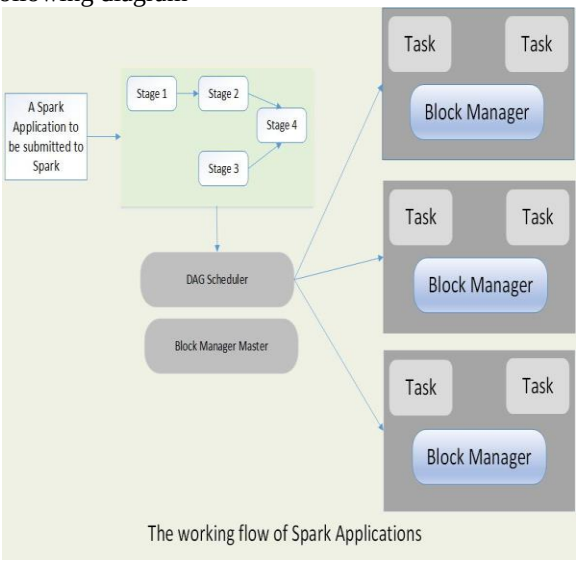

Fig.3: Working flow of Spark Application

An application is also called a driver on Spark. When a job is submitted to YARN, a runtime environment is created. At the same time a service called Block Manager is started on each node. Then the application is transformed to a DAG. The DAG scheduler is on duty of executing every stage of the process.

Some preparations are made before executing the driver program. YARN and spark environment needs to be deployed., and training data should be put on HDFS. Spark uses SparkContext to get access to a cluster. IP of master node, the name of the application, the SparkHome and the jar path to the SparkContext. SparkContext is used to get RDD and read files on HDFS. In driver program data is read from meta file to get the attributes and values put into HashMaps. The function textFile() is used to get the meta file on HDFS. After this, every line of RDD needs to be initialized.

The whole input data set is regarded as RDD. Cache () method is used to keep the RDD in memory for reusing. The function flatMap is used to get some lists of<key, value>. The reduceByKey function is used to get the sum needed to get the gain ratio of each node. When the information required is worked out, it is possible to get the attribute that has the highest gain ratio in each node on current layer

Initial Stage Preparation:

a) Configure Spark and YARN
b) Put the data into HDFS

Run C4.5 Tree Class (Driver Program)

SparkContext:

The constructor: new SparkContext (master, appName, [SparkHome], [jars]) is called to initialize SparkContext.

Initialization:Initial

Read and initialize attributes and their possible values from meta file

RDD

The input training set is regarded as a RDD on spark through textFile (path, missplits): RDD [String]

flatMap:

Get a list through each input line, including:

a) <id+att+value+class, 1>
b) <id,1>
c) <" total", 1>

Id means the unique number of a node on current layer

reduceByKey:

Get the sum of the same key from the RDD from flatMap.

The C4.5 code based on Spark is shownabove.

VI. METHODOLOGY

The flow of proposed system is as follows:

1) Configure Spark and Hadoop YARN environment.
2) Put the data into HDFS.
3) Load the customer data from HDFS as input for the algorithm.
4) Invoke the instance of C4.5 class.
5) When a job is submitted to YARN, a runtime environment is created and at the same time Block Manager is started on each node. Spark driver is transferred to DAG scheduler

6) Get the RDD and read files from HDFS.
7) Using flatMap and reduceByKey function the Spark framework get some lists of <key, value> and counts the no of occurrence of combination (id, attribute, value and class) and prints count against it.
8) Calculate entropy, information gain, SplitInfo and Gain ratio of attributes.
9) Process the input dataset from HDFS according to defined algorithm of C4.5.
10) Generate decision rules and store it in HDFS.
11) Accept the new test data from web UI.
12) Access the rules and based on it , decide the category of the new data
13) Provide the visualisation of the dataset from HDFS on the web UI in the form of bar graphics, pie charts etc. using Origin software

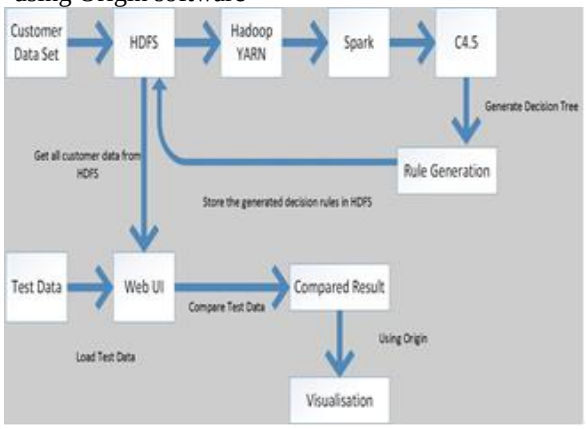

Fig.4: Working flow of the proposed model.

VII. RESULTS AND DISCUSSION

The existing system performs the customer behaviour analysis using Hadoop. The proposed system performs the customer behaviour analysis using Apache Spark over YARN. The proposed system runs application 100 times faster in memory and 10 times faster on disk than existing Hadoop system, as it requires only limited number of read/write cycle to disk and store intermediate data in memory. It is easy to program the proposed system as it has large number of high level operators with RDD. Spark is capable of performing batch, iterative and machine learning and streaming all in same cluster. As a result, it itself becomes a complete data analytics engine. The performance of the proposed system is much better than existing system which is understood from the following graph.

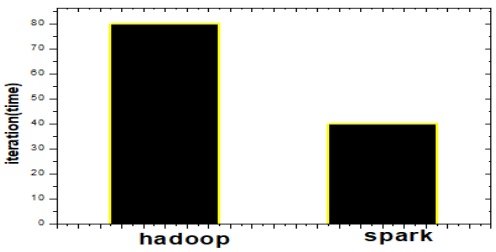

Fig.5: Performance difference between existing model and proposed model

The proposed system makes the accurate prediction of customers purchasing probability from the given dataset. The following graph depicts the prediction of purchasing probability across the months of a particular customer. The insight gained from the graph is visitors click is more on festive month – September.

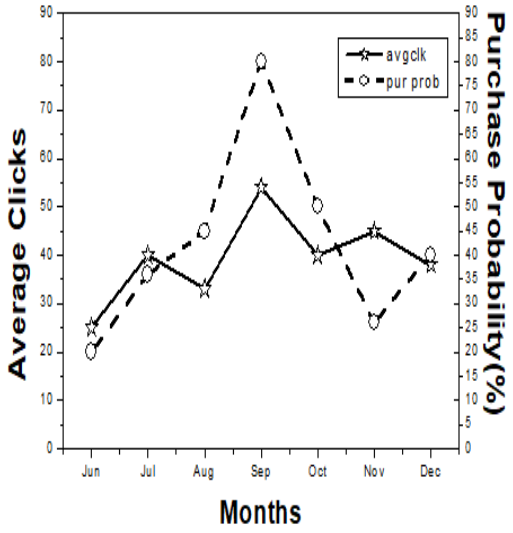

Fig.6: Average click and purchase probability across months

VIII. CONCLUSIONS

This paper defines the proposed model for distributed environment of C4.5 algorithm using Apache Spark over Hadoop YARN along with customer data visualization using Origin Software. With the advent and development of bigdata and cloud computing the traditional algorithm such as decision tree does not fit any more, hence introduced the implementation of statistical classifier C4.5 decision tree using Apache Spark over Hadoop YARN. Visualization is done using Origin which is very fast and user friendly. In future works we can use the new visualization tool such as D3.js which is reusable.

IX. ACKNOWLEDGMENT

I wish to express my deep sense of gratitude to **almighty, my teachers, parents and my husband** who lead me throughout my studies. I wish to express my gratitude to my guide **E K GIRISAN**, Sree Krishna Adithya College of Arts and Science, Coimbatore for his excellent guidance, encouragement and great involvement in my work.

X. REFERENCES

[1]. Laika Satish, Norazah Yusof. "A Review: Big Data Analytics for enhanced Customer Experiences with Crowd Sourcing." Procedia Computer Science 116(2017)274-283.
[2]. Hua Wang, Bin Wu, Shuai Yang, Bai Wang and Yang Liu. "Research of Decision Tree on YARN Using MapReduce and Spark."
[3]. Anindita A Khade. "Performing Customer Behavior Analysis using Big Data Analytics." Procedia Computer Science 79(2016) 986-992.

[4]. J. Ekanayake et al. "Twister: a runtime for iterative MapReduce, HPDC." 10 Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing Pages 810-818 (2010).
[5]. MapReduce, http://wiki.apache.org/hadoop/MapReduce.
[6]. YARN, http:// Hadoop.apache.org/.
[7]. Spark mllib, http://spark.apache.org.
[8]. J R Quinlan, "C4.5:programs for machine learning." Morgan Kaufmann 1993.
[9]. TomWhite, "Hadoop- The Definitive Guide" 3rd Edition, O'Reilly Media, Inc." Sebastopol, CA 95472, 2012
[10]. Eric T. Bradlow, Manish Gangwar, Praveen Kopalle, Sudhir Voleti. "The Role of Big Data and Predictive Analytics in Retailing" Journal of Retailing.
[11]. Agrawal.R and Srikant.R. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487-499 1993.b
[12]. Sheu, J.J., Chang, Y.W. and Chu, K.T., "Applying decision tree data mining for online group buying consumers' behaviour." International Journal of Electronic Customer Relationship Management, 2(2), pp.140-157 2008.
[13]. Romdhane, L.B., Fadhel, N. and Ayeb, B., "An efficient approach for building customer profiles from business data." Expert Systems with Applications, 37(2), pp.1573-1585 2010.
[14]. Caroline Lo, Dan Frankowski, Jure Leskovec, "Understanding Behaviors that Lead to Purchasing: A Case Study of Pinterest"
[15]. Bo Wu, Defu Zhang. "An Efficient Frequent Patterns Mining Algorithm Based on Apriori Algorithm and the FP-Tree Structure." International Research Gate of Computer Applications 108.16 (2010).
[16]. Alias Devi, P. Isakki, and S. P. Rajagopalan. "Analysis of customer behavior using clustering and association rules." International Journal of Computer Applications (0975–8887) Volume (2012).
[17]. Adhikari, Ratnadip, and R. K. Agrawal. "An introductory study on time series modeling and forecasting." arXiv preprint arXiv:1302.6613 (2013).
[18]. Dhandayudam, Prabha, and Ilango Krishnamurthi. "Customer behavior analysis using rough set approach." Journal of theoretical and applied electronic commerce research 8.2 (2013): 21-33.
[19]. Hussain, RZ Inamul, and S. K. Srivatsa. "A Study of Different Association Rule Mining Techniques." International Journal of Computer Applications 108.16 (2014).
[20]. Xurigan Saiyin, Chenna Hu, and Dou Tan1. "Research on Apparel Sales Forecast Based on ID3 Decision Tree Algorithm." arXiv ICMII 2015 (2015).
[21]. Xing, Shuning, et al. "Utility Pattern Mining Algorithm Based on Improved Utility Pattern Tree." Computational Intelligence and Design (ISCID), 2015 8th International Symposium on. Vol. 2. IEEE, 2015.
[22]. Pandya, Rutvija, and Jayati Pandya. "C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning." International Journal of Computer Applications 117.16 (2015): 18-21.
[23]. Joyce Jackson, Alva Erwin. " Combinatorial Temporal Closed+ High Utility Item set Mining Algorithm in Transactional Database" International Journal of Advanced Research in Computer and Communication Engineering.
[24]. Yuan, Baoli. "Applications and Analysis of Client Consumer Behavior Based on Big Data." Advanced Science and Technology Letters (2016).
[25]. Fokin, Dennis, and Joel Hagrot. "Constructing decision trees for user behavior prediction in the online consumer market." (2016).
[26]. Kalaivani Devi, T.Arunkumar. "Customer Behavior Predictive Modeling for Online Shopping using Tuned Decision Tree Method." International Journal of Computer Applications (01375–8887) Volume (2017).
[27]. Pawar, Namrata. "Analysis and prediction of E-customers' behavior by mining clickstream data using Naive Bayes." (2018).