Data Mining of Log Files Using Self-Organizing Map and Bisecting K-Means Clustering Methods Through Hadoop: A Review

Anita Choudhary[1] Mrs. Priyanka Dahiya[2]
[1,2]Department of Computer Science and Information Security
[1,2]Mody University of Science and Technology Lakshmangarh, India

Abstract— The continuous increase of computational strength has produced massive flow of data in past two decades. Big data is a data which cannot be processed and analysed by traditional techniques. It's not only used for store and handle large volumes of data but also to analyse and extract accurate information from the data in small amount of time. Today's internet world, data rapidly increases so analyse and storage becomes impossible and this also increases processing time and cost efficiency. In distributed computing various techniques and algorithm are used but problem remains still idle. To solve this problem Hadoop is used to process the files in parallel manner. E-commerce websites using log files analysing task to identify their user behaviour to improve their business. Large E-commerce websites like flipkart.com, amazon.in and e-bay.in millions of customers are visiting this sites simultaneously. As a result, these customers generate large amount of data in their log file entries. To analyse this large amount of log files entries we require parallel processing and reliable data storage system. In this paper, we present the Hadoop, bisected k-mean and SOM (Self-Organizing Map). Hadoop provides Hadoop distributed file system and MapReduce programming model to process huge amount of data in efficient and effective manner. Bisecting k-mean is used to analyse the existing offline data stream. Last method SOM is used to mine offline data streams using visualization tool like U-matrix methods.

Key words: Big Data, Hadoop; Web Log file; Data Stream; Bisecting K-means clustering; SOM and U-matrix, E-commerce

I. Introduction

Big data refers to technologies and architectures which are used to capture, store, process and run better quality volumes of data in small amount of time. Here, massive means data can range from petabytes to exabytes or to zettabytes. With the quick development of internet, e-commerce websites have brought large amount of data from their users. Behaviour information of the users are hide inside web logs. Web log is a log which is automatically created and maintained by the web server. Web log file contains details such as IP address of the computer which is making request, the date and time of the hit, the request method, the name and the location of the requested file, the HTTP status code, the size of the requested file etc. E-commerce websites mine the log files to predict the behaviour of the online users to provide product accommodation according to their previous purchasing products. Log file contain list of actions that occurred during when someone accessing the websites. Each user request are listed on a separate file in a log file this is called log entry. This entry is created every time when someone makes a request to your websites.

Log files are used to provide the pathway to complex systems so that whenever problem occur it's easy to find out and fix. But sometimes difficulty in reading the log files then log file analysis is necessary.

Hadoop framework provides reliable facility to store large log file in distributed manner and also provides parallel processing feature to process large log files in effective and efficient way. Hadoop divide the log files into multiple blocks and these blocks are evenly distributed across hundreds of nodes in a Hadoop cluster. It also replicates these blocks over multiple nodes to achieve reliability and fault tolerance. In one cluster two nodes are there: master node and slave node. Hadoop architecture includes two types of layers:

1) HDFS layer (Hadoop Distributed File System)
2) MapReduce layer

In Hadoop stack, MapReduce programming software is used to simplify the processing of large datasets, assign task by scheduling jobs for cluster node, guide the activities and re-execute the failure tasks. As the number of web user increases it's necessary to record user data in log file. Everyday large amount of data streams are generated so that we are using different methods to analyse data streams. Two types of data stream are there: offline and online

Web log files are offline type of stream. Most of the offline data stream mining uses k-means or hierarchical clustering algorithms. In clustering methods the data is divided into groups of similar objects. K-means clustering is better than hierarchical clustering when we are using large data sets but k-means also have disadvantages that resulting cluster depends on initial centroid. To solve this issue we are using Bisecting k-means. In this paper, we present the SOM (Self-Organizing Map) method to mine the web log files. This method used in web log mining for cluster analysis. SOM algorithm is selected for clustering that converts the high dimensional data into two dimensional maps by applying competitive learning techniques. This algorithm is used for solving the problems such as clustering and visualization.

Both SOM and bisected k-means methods are used to find out useful user request patterns so that website designed in dynamic manner using previous cluster patterns.

II. Web Data

In data mining task knowledge discovery is an important step for creating the suitable datasets. Web mining data is collected from server-side, client-side and proxy servers and organizations database. There are different kind of data available in data mining such as:

- Content: This refers to actual data in web pages.
- Structure: This type of data describes the content of organizational data. Intra page structure includes the arrangement of various HTML or XML tags. Main cause of using structure is hyperlinks are used to connect one page to another.
- Usage: This defines the pattern of usage of web pages like IP address, page reference, date and time of access.

- User profile: It provide demographics information about user which uses websites. It includes registration data and customer profile information.

III. Web Usage Mining

Web usage mining (WUM) is a division of web mining which apply the data mining techniques to find the user patterns. This appraise as business intelligence in an organization. Web usage mining is used to obtain the following things:

1) Understanding of user patterns.
2) To get the information by personalize the sites.
3) Construct tune up the server.
4) Revamp the site according to the preferences of user.
5) Construct business rules to get new clients and Marketing campaigns.
6) Construct efficient site.

Three main tasks for performing web usage mining which is shown in fig. 1

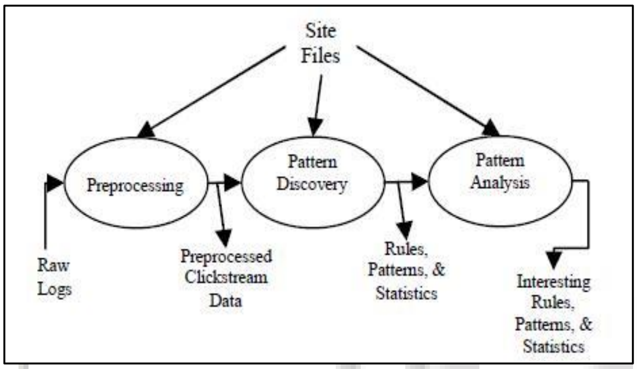
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, & Statistics
Interesting Rules, Patterns, & Statistics

Fig. 1: High Level Web Usage Mining Process [7]

A. Preprocessing:

Initially raw log data is inserted into pre-processing block. This involves two important steps which is User identification and session identification. In this three types of processing are done on log data:

1) Usage pre-processing: Usage pre-processing is used to include usage patterns of user.
2) Content pre-processing: It's used to access the content.
3) Structure pre-processing: It includes structure of the website.

This stage also involves data cleaning, efficient user identification, session identification, transaction identification and path completion. In data preprocessing change the format of the data to format which is convenient to user. In this filtering techniques are used to remove unwanted log entries from the log file using HTTP error code method.

B. Pattern discovery:

Output of the pre-processing stage that is pre-processed checksum data is taken as input and find out the specific user patterns to extract the useful information from log data. This patterns are obtained from different techniques such as statistical analysis, association rules, clustering, classification, sequential patterns and dependency modelling. Mostly statistical methods are used to analyse the user patterns.

C. Pattern Analysis:

In pattern analysis rules, patterns and statistics which is identified in pattern discovery step output is used as input to analyse the user pattern. It's used to identify the pattern accessible on page with a better level of consequence. This step consist of knowledge query mechanism such as SQL.

IV. Hadoop

To analyse the log file through Hadoop we consider a web application log file of banking server. Log file is simple text document in which various fields are there such as URL, date, hit, age, country, state and city. In this two phases are used to analyse the log file:

1) Pre-processing phase
2) Analysis phase

A. Pre-Processing Phase:

In this phase separation of fields using "#" symbol and remove the unwanted noisy data like multimedia files, style sheets etc. After pre- processing step log file is stored in HDFS.

B. Analysis Phase:

Log files is distributed over Hadoop cluster using MapReduce algorithm and execute the pig query to integrate and categorize the results.

1) Distribution of Log Files:

Hadoop divide the input file into smaller blocks which are equal in size and distribute these block into multiple nodes in Hadoop cluster. MapReduce is used to process these blocks in parallel way.

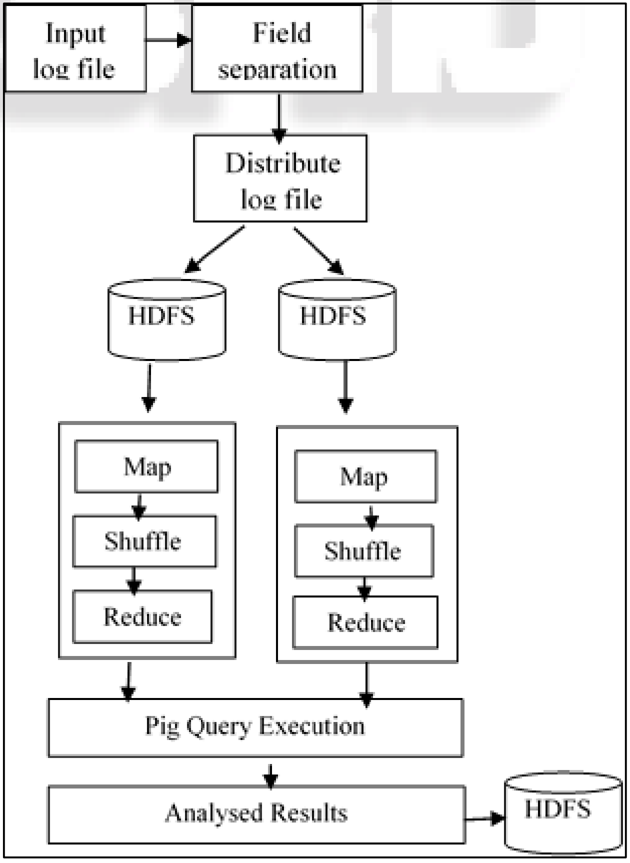

Fig. 2: Distribution of log files

2) *MapReduce Algorithm:*

It's a programming model which is easily scalable over multiple nodes in Hadoop cluster. Java programming is used to execute map and reduce functions. MapReduce takes log file as input and feeds all the records of the log files into mapper. Mapper is used to process all the records of log files and reduces processes all the outputs of mapper and produce final results. Mapper is also used to filter and transform into which form so that reducer can easily aggregate and recognize the form and we get appropriate output from it.

a) Map Function:

In Map function InputSplit of log file take as input. It produces the result in form of pairs (key, value). Each time when key is occurred it emits (key, '1') pair. If appearance of key is n, then it produces n (key, '1') pairs. Mapper provide the OutputCollector utility to collect the output from mapper and reducer. Reporter utility is provided to report the progress of application.

```
Map (LongWritable Key, Text value, OutputCollector output, Reporter reporter)
{
        For each key in the value;
        EmitIntermediate (key, '1');
}
```

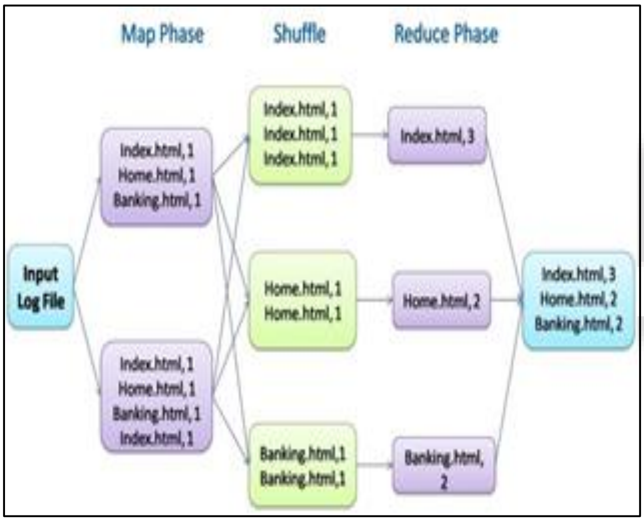

Fig. 3: MapReduce Framework [6]

b) Reduce Function:

The output of the mapper function (key, value) pairs is the input of the reducer function. This function is used to sums together all the counts values which is emitted by mapper function. If input value of reduce function (key, (1, 1, 1….ntimes)) then this function combine all the values for that key producing output (key, n) pair. OutputCollector and reporter work similar to mapper method.

```
Reduce (Text key, Iterator values, OutputCollector output, Reporter reporter)
{
        int sum = 0;
        for each v in values;
        sum+= ParseInt(v);
        output.collect (key, (sum));
}
```

3) *Pig Query:*

This type of query uses Pig Latin language. Pig Latin statements are arranged in the following manner:

1) LOAD statement: This is used to read data from Hadoop file system.
2) Transformation statement: Series of this statements are used to process the data.
3) STORE statement: This type of statement is used to write output to the Hadoop file system.

Example of Pig Latin is show in below which uses results of MapReduce to get results of total hits per pages:

```
A = load '/home/hadoop/output/1/part-00000' using PigStorage ('\t') AS (page:chararray,hits:int);
B = load '/home/hadoop/output/2/part-00000' using PigStorage ('\t') AS (page:chararray,hits:int);
X = UNION A, B
Y = FILTER X BY (page matches '^HitsPage-.*')
; X = FOREACH Y GENERATE page, hits;
X =GROUP X by page;
X = FOREACH X GENERATE group, SUM(X.hits);
store X into 'Data/HitsPages' using PigStorage('\t','-schema');
```

V. Bisecting K-means

Bisecting k-means is a collaboration of k-means and hierarchical clustering. We combine these two algorithm to use advantages of both so that we can get accurate results when we are using large number of datasets. In simple k-means method we partitioning the data in k clusters centers to find out the user patterns but in Bisection method split one cluster into two sub clusters at every bisection method using k-means method. We divide the clusters till final k-clusters are not obtained.

Bisecting k-means based on simple k-means method and it uses merits of k-means and it also have advantages over k-means.

Advantages of bisection k-mean method:

1) Bisecting k-means is more effective as compare to k-means method.
2) In k-means every data computation point of the data set and k centroids are used but in bisection method only the data point of one cluster and two centroids used in computation. Thus bisection method reduced the computation time.
3) Bisection k-means produce same size of clusters but k-means is known to produce different size of clusters.

Steps of bisecting k-means algo for finding k clusters: [9]

1) Select a cluster to split.
2) Using basic k-means algorithm finding the 2 sub clusters. (Bisecting step)
3) Repeat the bisecting step, for ITER times and take split part to generate the clustering with the highest overall similarity.
4) Repeat steps 1, 2 and 3 till the desired number of clusters is reached.

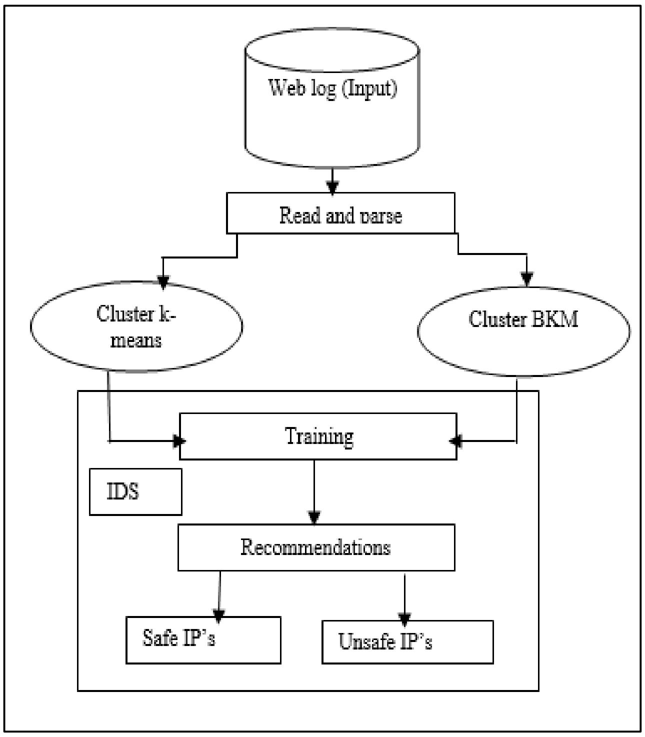

Fig. 4: Flow of design modules

There are different ways available to choose for splitting the cluster. For example if we choose largest cluster at each step, one with the least overall similarity or we can choose cluster using a criterion based on size and overall similarity.

Web log data which is in pdf are collected from college website in which various reports and summaries available. These files are in non-standard format. Firstly this files converted into standard format in which extraction of information is done. Extraction procedure includes taking information about account bandwidth and web usage reports and summaries. Pdf files are firstly read then parsed. Here parsing means analysing the text and find out the useful information from them. Parsing consists of displaying IP address from the bandwidth reports and the total bytes communicated with it. Bisection k-means clustering is done in order to get identical IP address and packet combinations together. We design a framework which will be able to detect the intrusion which are present in the network environment. Training is important in bisection method because every recommendation system check the input signatures (IP address, cluster no., packet size etc) are proper or not. Training part is used to create a database having number of values that contain IP signatures. Comparison of IP address occurrence with the IP address of the same signatures in the database. We find the best matched signature which is given as a class of the input IP. If IP address and cluster are matched then the difference in packet size is calculated as packet size of matching entry. If the difference in sizes are minimum then we use this packet to decide whether the particular IP as considered as safe IP or as an infected IP. When counts of each class is done then recommendation output is generated. In this each IP is checked to find the values of its signature.

VI. SELF-ORGANIZATION MAP

Self-organising map is unsupervised approach which is typical unsupervised learning neural networks used to project a high dimensional input space on a low dimensional topology so as to clusters directly. Teuvo Kohonen introduced a SOM network used to reduce the dimension of data through using self-organizing neural networks. SOM network build a map of one or two dimensions which is used to plot the similarities of data by grouping the same type of data items together. This type of mapping process reduces the dimension problem. SOM network combines the dimension reducing and clustering in one network and organized the documents into meaningful clusters.

It's used in mining of log data and clustering the data without knowing the class. Web pages with the same patterns are clustered together using association rules. The cluster which is generated by SOM has specific meaning to web browsing behaviour. Thus the SOM can treat as cluster analysing tool of high dimensional data.

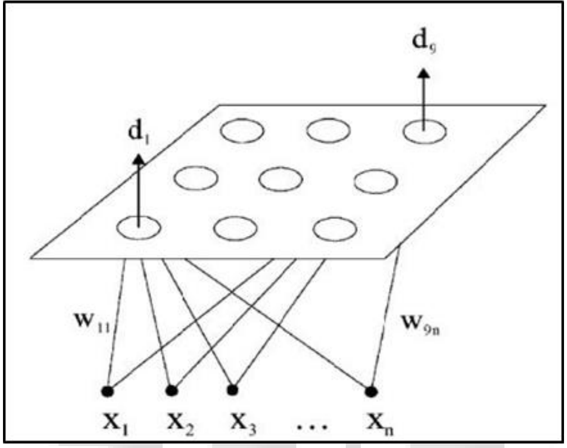

Fig. 5: The mapping from a one dimensional input to two dimensional array [7]

This figure shows the mapping of one dimensional input to two dimensional array by using the visualization tool that reduces the dimensions using the competitive learning techniques. The SOM network assemble itself by competing description of the samples.

Neurons are also permit to alter themselves in hoping to win the next competition. This selection and learning process makes the weights to arrange themselves into a map representing similarities.

The steps which are involved in SOM algorithm is as follows. [12]

```
1) Initialize Map
2) Set t = 0 and repeat the following steps until t > 1
       Arbitrary select a sample
       Get best matching unit
       Scale neighbors
       Increase t by a small amount
3) End for
```

In SOM algo we firstly initialize the weight vectors. Then selects the sample vector arbitrary and searches the map of weight vectors to discover the weight that can represent the sample best. Every weight vector has a location, it also has neighbouring weights that are close to it. The chosen weight performance is better than arbitrary selected sample vector. This step we increase t in small amount because the number of neighbours and how much every weight can learn decreases over the time. The advantage of SOM, it automatically clusters the documents. It can also be applied to large scale of data.

VII. SOM BASED WEB PAGE CLUSTERING

This approach is divided into three steps: data pre-processing, Web page mapping and clustering analysis. In pre-processing step pair of methods are used to identify users, transactions and sessions. In processing step web site topology is also identified to filter out the user patterns. Web data is pre-processed in this step for further processing. After that SOM is used to cluster pages from similar navigating patterns. SOM method uses current user navigation pattern. SOM uses k-means clustering in which one or more clustered are selected for further analysis. In clustering analysis step, result of web page mapping are stored in two dimensional array. This analysis is used to find out the user patterns and predict the users move when they browsing some particular sites.

VIII. CONCLUSION

This paper described the big data in all possible aspects. Big data refers to large amount of data which can't be handled by traditional systems and takes lot of time for processing. Therefore, this paper focused on distributed approach that's hadoop to get the results in close real time. Hadoop uses MapReduce programming to process structured, semi-structured and unstructured data sets. Hadoop cluster is used to extract the useful information from log files which is in the form of text file and log file takes lots of time to process when we are using traditional systems. In this paper clustering methods are used to detect the intrusion which are present in network environment. It also helps to classify the IP's in safe IP and unsafe IP's. Two types of clustering are used to mine the log data: k-means and bisected k-means. In this paper we are using bisected method to generate uniform clusters and it can't be generate the empty cluster which are possible in simple k-means method. It takes less amount of time with accuracy and more efficient when number of clusters are greater. Bisecting k-means is dependent on initial number of cluster 'k' which is define by user. We combine the both bisecting k-means and SOM to extract the valuable user request patterns so that web site dynamically designs according to past cluster patterns. SOM Uses U-matrix for visualization and k-means uses probability distribution for that.

REFERENCES

[1] Bina Kotiyal, Ankit Kumar, Bhasker Pant and RH Goudar, "Big Data: Mining of Log File through Hadoop", In: Human Computer Interactions (ICHCI), 2013 International Conference on Big Data.

[2] T. Nadana Ravishankar and Dr. R. Shriram, "Mining Web Log Files Using Self-Organizing Map and K-means Clustering Methods", International Journal of Applied Engineering Research, 2015.

[3] S Saravanan and B Uma Maheswari, "Analysing Large Web Log Files in a Hadoop Distributed Cluster Environment", International Journal Computer technology and Applications (IJCTA), vol. 5, 2014.

[4] Savitha K and Vijaya MS, "An Efficient Analysis of Web Server Log Files for Session Identification using Hadoop MapReduce", Elsevier, 2014.

[5] Dipali Patil, Snehal Patil and Payal Balsetwar, "Excerption of User Profile from Web Log Data using Hadoop Framework", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 4, April 2013.

[6] Sayalee Narkhede and Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs over Hadoop MapReduce", International Journal of UbiComp (IJU), Vol.4, No.3, July 2013.

[7] Hemanshu Rana and Mayank Patel, "A Study of Web Log Analysis Using Clustering Techniques", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 1, Issue 4, June 2013.

[8] Ruchika R. Patil and Amreen Khan, "Bisecting K-Means for Clustering Web Log data", International Journal of Computer Applications, Volume 116 – No. 19, April 2015.

[9] K. Poongothai, M.Parimala and Dr. S. Sathiyabama," *Efficient Web Usage Mining with Clustering*", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 6, No 3, November 2011.

[10] Paola Britos, Damián Martinelli, Hernán Merlino and Ramón García-Martínez, "Web Usage Mining Using Self Organized Maps", International Journal of Computer Science and Network Security (IJCSNS), VOL.7 No.6, June 2007.

[11] P.Nithya and Dr. P.Sumathi, "A Survey on Web Usage Mining: Theory and Applications", International Journal Computer technology and Applications (IJCTA), Vol 3 (4), 2012.

[12] A.M.Sote and S.R.Pande, "Web Page Clustering using Self-Organizing Map", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 4, Issue. 1, January 2015.

[13] Peilin Shi, "An Efficient Approach for Clustering Web Access Patterns from Web Logs", International Journal of Advanced Science and Technology, 2009.

[14] K. Poongothai, M.Parimala and Dr. S.Sathiyabama," Efficient Web Usage Mining with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[15] L. Jing-min and H. Guo-hui, "Research of Distributed Database System Based on Hadoop", IEEE International conference on Information Science and Engineering (ICISE), pp. 1417-1420, 2010.

[16] Dehu Qi and Chung-Chih Li, "Self-Organizing Map based Web Pages Clustering using Web Logs", Proceedings of 16th International Conference on Software Engineering and Data Engineering, SEDE, Las Vegas, Nevada, Pages 265-270, July, 2007.

[17] Supinder Singh and Sukhpreet Kaur, "Web Log File Data Clustering Using K-Means and Decision Tree", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 8, August 2013.