International Journal For Technological Research In Engineering
Volume 2, Issue 10, June-2015 *ISSN (Online): 2347 - 4718*

ENHANCING CLUSTERING PERFORMANCE USING MODERATE NUMBER OF NODES

Tanbir Singh Khaira[1], Priyank Singh Hada[2], Surinder Kaur[3], Amaan Imam[4], Aishwarya Shekhar[5]
[1,2,4,5] School of Computing and Information Technology, Manipal University Jaipur, India
[3]Punjab Technical University

Abstract: We are very aware of the amount of information around, this is data the age of enormous amount of data. So enormous is the amount at which this data is being generated all around from various fields. This data varies to petabytes and even beyond this limit. This is a huge challenge to store analyze data for all the important purposes, this brings in the concept of data mining along with many other derived from this intelligent field. Clustering of the data is the sub field in the data mining. Clustering is the unsupervised process of separating a set of data elements into meaningful groups known as clusters. It is known as unsupervised learning because datasets are assigned to a cluster without knowing to which this dataset belong or there are no predefined classes available. It is a very common technique for data analysis, data which is static and it is used in many fields like big data, machine learning, image analysis etc. basically they usually identify the natural clusters of data collection. K-means and Apriori is a good example of such method.
Keywords: Apriori, Clustering, Data Mining, K-means.

I. INTRODUCTION

We are living in the era of data age. The size of data is expanding day by day for collection and also for retrieval. It is not easy to maintain large volume of data sets electronically but now days there will be good methods have been generated through which we can store and access the data very easily. There are some figures for data count per minute which is increasing from kilobytes to megabytes, gigabytes to terabytes and so on. There is research going on nearest neighbor, Apriori for search in highest dimensional spaces. By this process data is to be stored in cluster form multiple source and then segmented into groups which is stored on various disks. This created a indexing for accessing the file from stored place and the processing speed is increased for query [1]. So there's a lot of data out there. But you are probably wondering how it affects you. Most of the data is locked up in the largest web properties (like search engines) or in scientific or financial institutions.The problem is simple: although the storage capacities of hard drives have increased massively over the years, access speeds, the rate at which data can be read from drives have not kept up. One typical drive from 1990 could store 1,370 MB of data and had a transfer speed of 4.4 MB/s so you could read all the data from a full drive in around five minutes [2]. Over 20 years later, one terabyte drives are the norm, but the transfer speed is around 100 MB/s, so it takes more than two and a half hours to read all the data off the disk. This is a long time to read all data on a single drive and writing is even slower. The obvious way to reduce the time is to read from multiple disks at once. Imagine if we had 100 drives, each holding one hundredth of the data. Working in parallel enables the reading of the data much faster up to less than two minutes [3]. Using only one hundredth of a disk may seem wasteful. But we can store one hundred datasets, each of which is one terabyte, and provide shared access to them. We can imagine that the users of such a system would be happy to share access in return for shorter analysis times, and, statistically, that their analysis jobs would be likely to be spread over time, so they wouldn't interfere with each other too much. So there several indexing methods proposed for resolving this problem such as NV-TREE and LSH and Clustering etc. By creating this kind of index which helps in querying the data from collection, comparison if data points. By using good technique of clustering the disk reading is minimizing. When we were working with large data then, indexing helps in disk reads from multiple data source of data collection at an average time. Large amount of data was collected which leads to rich data but poor information situation. The problem is that the decisions are not based on rich information data that are stored on different data repositories rather they are based on decision makers intuitions because they don't have the appropriate tools to extract the knowledge embed in various data repositories. Data mining is the process of mining knowledge from the data. Intelligent methods were applied to extract the data patterns from massive amount of data.

II. CLUSTERING

It is a very common technique for data analysis, data which is static and It is used in many fields like big data, machine learning, image analysis etc. basically they usually identify the natural clusters of data collection. K-means and Apriori is a good example of such method. K means attempt to find natural clusters by recursively. Recursively clustering is a process of searching the natural clusters from the data collections to make it more convergence, repetition is done just for convergence. Chierichetti et al. proposed a method called "cluster pruning" [4]. Cluster pruning consists of selecting random cluster leaders from a set of points, and the rest of the points are then clustered based on which cluster leader they are closest to. Clustering is the unsupervised process of separating a set of data elements into meaningful groups known as clusters. It is known as unsupervised learning because datasets are assigned to a cluster without knowing to which this dataset belong or there are no

predefined classes available [5]. The better quality clusters will be produced by the clustering algorithms which satisfy the following conditions:

- The Similarity factor within same cluster must be high which is also known as intra cluster similarity.
- The similarity factor between different clusters must be low which is also known as inter -cluster similarity.

III. TYPES OF CLUSTERING ALGORITHM

- Partioning based: Build various partions and then evaluate them based on some criteria.
- K-Means: Every cluster is represented by center of the cluster.
- K-Medoids: Every cluster is represented by one of the object in the cluster.
- Hierarchal based: Create a hierarchal breakdown of data objects.
- Agglomerative: Start with single cluster and at each step join two closet clusters.
- Deglomerative: Start with one cluster and then divide that cluster into sub clusters and progress recursively on each sub set.
- Sting: A Statistical information grid approach for spatial data mining.
- Clique: Used to cluster high dimensional data stored in large tables.
- Dbscan: it is density based algorithm creates cluster based on density distribution of agreeing nodes.
- Optics: density based clustering algorithm which finds clusters in spatial data.

IV. DATA MINING ISSUES

A. Mining methodology: Various kinds of data can be mined including mining different kinds of knowledge, knowledge from different disciplines, handling incomplete and noisy data and mining of data in different formats [6].

B. User interaction issues: These issues include interactive mining at multiple levels of abstraction, assimilation of background knowledge, Knowledge of data mining query languages and ad hoc data mining is required, user friendly interfaces are required to present the data in different formats [7].

C. Performance issues: efficiency and scalability of data mining algorithms, parallel, distributed and incremental algorithms.

D. Database diversity issues: handling of relational and complex data, heterogeneous databases and global information systems.

E. Impact of data mining on society: The first issue is what type of technology is benefitted to the society, How to ensure the privacy of data that can be mined and the last and the most important is issue of invisible data mining which means mining of data without having the information of mining algorithms [8].

Fig 1. Data Mining Issues

V. METHODOLOGY

Our aim is to make our technique speedy and reduced time complexity in all respect of collecting data. Our proposed clustering technique is very useful for the stability and robust point of view for this scenario. According to K-mean clustering algorithm we could not creates a index for large data sets in limited number of time [9]. Here, we proposed an algorithm which can search and indexed the data at very high speed with low latency. When we focused on the previous algorithms they all are very slow in processing the data, which affects the different parameters of clustering algorithm like, buffer capacity, Agility, Multi-tenancy, Peak-load capacity, reliability and Utilization and efficiency [10]. In our proposed algorithm we are making the cluster heads from data which increases the processing speed.The following algorithm is divided into 3 steps .The first step scans the database, searching algorithm starts working for finding the cluster heads to form the clustered data[11]. Cluster Heads are identified on the basis of their attributes and cardinality. So, we are focusing on performance and significance of cluster heads which means higher the significance of cluster heads is directly proportional to higher its suitability for clustering ensemble. In this phase scanning of database for categorical data has been done which can improve the efficiency because of the extraction of clusters from the categorical data. Finally the large data set will be generated which will be output[12].

VI. ALGORITHM

```
Begin
//Initialization Step
        Support=20%
        If(transactionID > Support)
                Select item in mining frequent item set in
DB
//Transformation Step
        For k=i to n //mining in vertical data
                If(transactionID <support)
```


Learn item sets
//Final Reduction
Total result & o/p
End

VII. RESULTS

The input files are uploaded to eclipse, which check the files for data mining according to the replication level. The input consists of text files with multiple methods descriptors per line, where each descriptor has 128 dimensions and parameters. All dimensions in each descriptor must be converted from a string to a byte (char) representing that number. This has some impact on the processing time; better performance can be obtained by having the input data as text. Running the original cluster pruning with text input was three times faster than running it on image input. All programming of the cluster pruning method was done in object oriented programming and it was used to execute the java code within the reduce java framework. Many tools was used for memory manipulation, among other features, and Threading Building Blocks (TBB) for local parallelism.

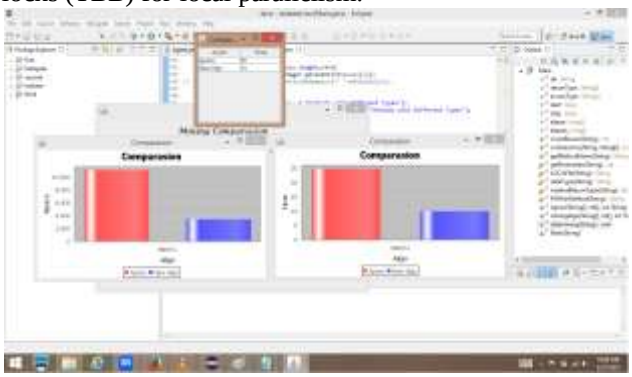

Fig.2 Result of The Proposed Algorithm

VIII. CONCLUSION & FUTURE WORK

The proposed work was the clustering algorithm, where each searching and classification of data is done. It is done in the way where worker creates its own index and clustered file, and both files are stored on its local storage. With this method each local index is smaller, but each query descriptor must be replicated to all workers During search and formation of clustered data is done. By this proposed work the processing time and space is reduced to 60 %. Hence it would be more beneficial for future if we implement this in Hadoop environment. Future Scope: Some extensions to the Hadoop framework are needed for implementation, however, so that it can process binary files. A custom File Input Format, File Output Format, record reader and record writer are needed for Hadoop to process binary data. This is an interesting direction for future work.

REFERENCES

[1] Tian Zhang,Raghu Ramakrishnan,Miron Livny“ BIRCH ,2013.
[2] Joshi Aastha, Kaur Rajneet “ Comparative Study of VariousClustering Techniques in Data Mining “ International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X,2013.
[3] Shreya Jain, Samta Gajbhiye “Comparing and Selecting Appropriate Measuring Parameters for K-means Clustering Technique” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307,2012.
[4] Jagadeeswaran V.S., P.uma “Detection of noise by efficient hierarchical birch algorithm for large data sets” International Journal of Advanced Research in Computer and Communication Engineering, ISSN 2319-5940,2013.
[5] Jagadeeswaran V.S., P.uma “Detection of noise by efficient hierarchical birch algorithm for large data sets” International Journal of Advanced Research in Computer and Communication Engineering, ISSN 2319-5940,2013.
[6] P. Indira Priya, and Dr. Ghosh D.K. “A Survey on Different Clustering Algorithms in Data Mining Technique” International Journal of Modern Engineering Research (IJMER) 267-274 ISSN: 2249-6645 ,2014.
[7] Shreya Jain, Samta Gajbhiye “Comparing and Selecting Appropriate Measuring Parameters for K-means Clustering Technique” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307,2012.
[8] Shraddha Shukla and Naganna S.“A Review ON Kmeans DATA Clustering APPROACH”International Journal of Information & Computation Technology.ISSN 0974-2239,2014.
[9] Sharma Narendra, Bajpai Aman, Mr. Litoriya Ratnesh “Comparison the various clustering algorithms of wekaTools” International Journal of Modern Engineering Research ,ISSN 2250-2459,2012
[10] Er. Gupta Arpit,Er. Gupta Ankit ,Er. Mishra Amit “Research paper on cluster techniques of data variations” International Journal of Advance Technology & Engineering Research (IJATER) ISSN NO: 2250-3536,2012
[11] Jain Anoop, Bajpai Aruna, Rohila Manish Kumar, “Efficient Clustering Technique for Information Retrieval in Data Mining “International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459,2012.
[12] Tian Zhang,Raghu Ramakrishnan,Miron Livny“ BIRCH : A new Data Clustering Algorithm and its Applications”, Data Mining and knowledge Discovery.Volume1,Issue2 pp141- 182,1997.

