Distributed Data Clustering

Sundas Charagh[#1], Ayesha Saleem[*2], Amnah Mukhtar[#3]

[#]*Students of MS(cs) Department of Computer science University of agriculture Faisalabad Pakistan*

***Abstract*—** *In modern era the volume of data is enlarging day by day. It has become impossible to handle this data without data mining there are different techniques-clustering is one of them. Clustering is a process of grouping same type of objects. In distributed data clustering these groups are distributed over different sites and then centralized at global sit. The purpose of distributing these clusters is efficiency, performance, communication cost and storage limit. There are many different techniques and algorithms are available for distributed data clustering. These algorithms are divided into two categories-synchronous and asynchronous that further has some sub-categories such as k-means, k harmonic means, DBSCAN, PCA based and many more. The paper also describes some important merits of distributed data clustering as well as de-merits.*

Keywords- *Data mining, Clustering, efficiency and performance.*

I. INTRODUCTION

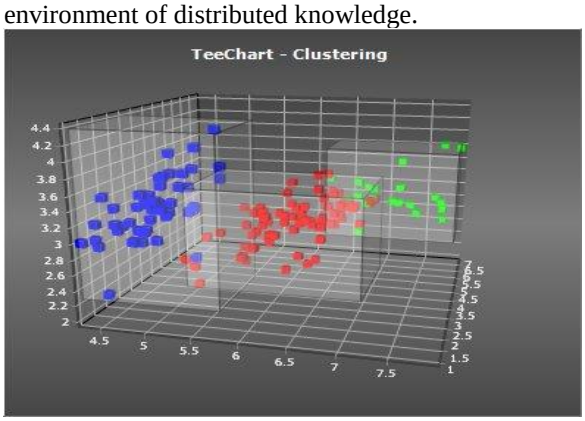
In a decade, the degree of data in our organization escalating day by day, it is due to numerous factors like the computerization of the data attainment and condensed cost of storage. It is impossible to handle large amount of data without data mining. Moreover, a emergent interests in origination the techniques of recorded data and the computational algorithms which are used to extract relevant information. In data mining, we apply various techniques and methods to databases, while having an objective to mine hidden information in huge bulk of data. Data clustering or analysis of clusters is a job of combining a complete set of things or objects in a manner that objects may relate with the similar groups (known as a cluster) are more alike to each, excepting to those that are in some other group (are clusters).It's a key operation of exploratory mining of data, and also a ordinary method of statistical analysis of data, which can be used in a lot of fields, together with pattern recognition, information retrieval machine learning, image analysis, and bioinformatics [4].

Data Mining

It is a method of mining useful data and patterns from huge databases. It is being used in a number of distinct regions of science and business, like banking, biology, insurances, medicine. There are several problems of data mining problems also exist such as classification, discovering of association rules, segmentation clustering. Knowledge Data Discovery can be described as a widely used multi stage procedure, interactive non-trivial and iterative, to recognize comprehensible, novel, valid, usefully, ultimately, potentially and understandable patterns. The amount of existing data is massive (it is considered in gigabytes and terabytes), consequently it is essential for user to pertain high performance of algorithms etc using distributed and parallel computing. The Distributed Data Mining aims to resolve problems of data mining by facilitating latest distributed algorithms that are appropriate to clustering environment.

Distributed data clustering

Owing to advancement in infrastructure and computation of technology, the numeral distributed information supplies are easily reachable to a user because resources have grown-up rapidly. For effective usability of distributed information, it is pleasing to permit collaboration and coordination among a variety of sources. It is particularly relevant to community administration as an instance. But here are limitations because the new society is flattering more and more reliant on the information day by day as well as constraints are increasing such as an individual constraint of corporate confidentiality and privacy are arising. But the solution is hidden in data mining because clustering is an approach that offers us an opportunity to deal and encounter with different types of constraints efficiently and in appropriate manner. It also permits to classify objects and attributes (customers) and the understanding of correlation sources such as services given by organizations where objects may originate from each other. Here the term co-relation means for collaboration.[3] Each site of distributed environment provides any type of information about an object or entity such as customer including the tasks performed by central or global site (may any type of data) it is done to obtain an advantage from distributed environment of distributed knowledge.

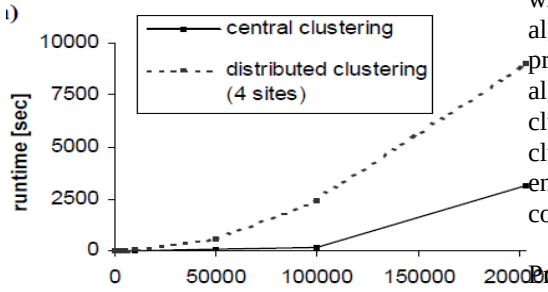


In distributed clustering, our objects that have to be clustered are on different sites. as a replacement

of transmitting all entities and objects to central site (also designate as a server) due to it we are concerned with customary clustering algorithms to explore the data, on the local sites the clusters of data are independent called clients. In a following step, the main central site makes an effort to set up a centralized clustering which is based on a limited model. Clusters of distributed environment lead us to two varied levels- global and local level

Local and global site

In local level, all sites have independent clusters, self-managed and interact with each other independently. When we have completed the clustering, the local site reflects a best optimal trade-off between accuracy and complexity. Our projected local level comprises of a group of representatives for every local cluster. Each ambassador is an object that is concrete which is stored on local site. While building local clusters, involvement of each independent and individual cluster of data is in use as a complete. To deal with constraints, the solution is to communicate with the data sets after contract on the privacy connecting both of distributed sites. When it is collected from various sites we need to have formatted data to come about with an individual data cluster. We should have to take care of data at central site because there are many issues regarding to structure.[1][2] Data is simply merged collectively due to structure resultant data cluster comprises of upcoming features of individual and independent data set.



From distributed model, we need data for beginning process which is needed to compute the output of global cluster with the contribution of independent data sets at the central site, there are many clustering approaches Hierarchy algorithms, Partitioning algorithms, Grid-based, Density-based, Model-based. Because of cost communication, privacy and storage limit, distributed data clustering functional when we try to centralize distributed data[5].

Distributed Data clustering algorithms

When we use distributed data clustering approach various algorithms are needed to deal with the requirement of distributed data. There are two type of clustering algorithms.

The first type of clustering algorithm methods is of message passing through multiple rounds. This require synchronization amount of significant.

In a second type of clustering algorithm methods is to build local model of clustering then pass it to the central which is known as asynchronous. The global model is a combination of central site. Single message passing round are required by this method to fulfil optimum synchronization requirements.

The related algorithms are:

Synchronous algorithm:

- K-means for inherent data parallelism
- K-harmonic means for inherently homogeneously distributed data
- clustering technique based on PCA for heterogeneous distributed data
- clustering technique based on Kernel-density for heterogeneous distributed data

Asynchronous algorithm:

- Hierarchical Distributed clustering algorithm which is based on heterogeneous distributed data
- Spatial clustering for homogeneously distributed data
- Merging hierarchical clustering for homogeneously distributed real-valued data
- DBDC for density based clustering algorithm require large memory space use globally and locally perform DBSCAN

These algorithms are designed to guaranty accuracy which is optimal for the required solution. These algorithms have the ability to deal with various problematic scenarios but the selection of the algorithm is very crucial element to achieve the clustering requirement which is as important as in clustering environment as well as distributed environment with little communication fraction for collecting all the data from single location[7]

Privacy

Privacy plays a vital and significant role in DDM, some contributor don't desire to carve up their data but they have to do for some reasons to take part in DDM. Security and privacy gives one more motivation of distributed KD. There are many constraints and integrity constraints on privacy to avoid unauthorized access of any data. These are very beneficial for global data mining. For example a corporation clusters its client's data to restart marketing campaign. Keep in mind, it a multidimensional and multinational corporation.[2]

The three privacy constraints regarding to it are

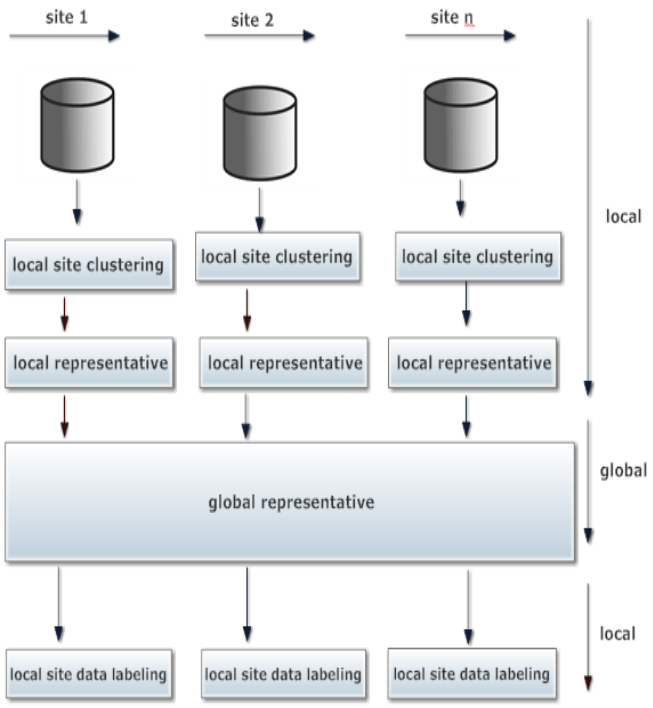
- What we want to mine? Here we first determine from which type of architecture we are using.
- How our data is distributed? Either is horizontal or vertical?
- What are the security constraints that we have to fulfill to achieve our requirement?

Some Privacy issues related to data mining are data distortion and the sanitation approaches.

Sanitation: sensitive data pattern is a serious problem because it is a crucial task. So we have o modify it in some pattern that can be mined. These are developed to handle privacy and integration rules. Different techniques are developed by many researchers. The basic idea is to modify or remove items in database to minimize or enhance the support of mostly used data items or sets, through it data owner are able to conceal sensitive patterns or frequently used items with a little impact on non-sensitive data sets

Data distortion: This approach provides privacy for single record of data through a little modification of in original data. These approaches intend to design many distortion methods. After it the ascertain of real values of any record. On the global end, properties remain unchanged. for example data distortion is applied to the classification of decisions based on tree. The weaknesses of data distortion are data privacy under some condition, reconstruction of original data with probability [2][6]

Relationship of local and higher cluster



Advantages of distributed data clustering

- Distributed data mining methods can offer saving in a processing time through use of inherent parallelism in a distributed system, storage cost because the data does not need to be copied and the human cost of integrating data into a ware house.
- Increase Scalability by allowing you seamlessly add new component as your business growth requires
- High dimensionality
- Privacy and security concern provide another motive for distributed knowledge discovery, data is distributed because it has been collected are produce by different parties. Privacy can release of the data in global data mining.
- Interoperability and usability
- Simplified administration of server by allowing you to manage a group of system as a single system or as a single database
- Minimal requirements for domain knowledge to determine input parameter
- Recovery availability is easy by using distributed data clustering
- Ability to deal with noisy data missing erroneous data some algorithms are sensitive such data may lead to poor quality cluster
- Distributed query processing

Disadvantages of distributed data clustering

- In distributed data clustering large volume of the data must be transfer to the centralized server which is expensive in term of communication cost.
- Dealing with large number of dimension and large number of data items can be problematic due to time complexity
- Reliance on prior knowledge and user defined parameter
- Data order dependency
- Data integration is a big issue in a distributed data clustering
- Many clustering techniques are based on trying to minimize or maximize a global objective function. The clustering problem then becomes an optimization problem.
- Algorithm dependency-without choosing efficient algorithm leads to poor performance

Conclusion and future work

The paper presents an overview of data distributed clustering that distributed data clustering is an important technique of data mining that helps you to deal with large amount of data. Clustering is a collaboration of local and global model for ease of availability and access in minimum cost with better performance. There are many open issues are also available regarding to privacy, performance and efficiency. We should investigate them very carefully in future.

REFERENCES

[1] V. Fiolet, E. Laskowski, R. Olejnik, L. Ma, B. Toursel, and M. Tudruj, "Optimizing Distributed Data Mining Applications Based on Object Clustering Methods," pp. 1–6, 2006.

[2] X. Lin, C. Clifton, and M. Zhu, "Privacy-preserving clustering with distributed EM mixture modeling," *Knowledge and Information Systems*, vol. 8, no. 1, pp. 68–81, Dec. 2004.

[3] H. Kriegel, "Towards Effective and Efficient Distributed Clustering," 2003.

[4] J. C. Silva, C. Giannella, R. Bhargava, H. Kargupta, and M. Klusch, "Distributed Data Mining and Agents f g f g," no. Ddm.

[5] E. Januzaj, H. Kriegel, and M. Pfeifle, "DBDC : Density Based Distributed Clustering."

[6] I. S. Dhillon and D. S. Modha, "A Data-Clustering Algorithm On Distributed Memory Multiprocessors."s

[7] Forman, G., & Zhang, B. (2000). Distributed data clustering can be efficient and exact. *ACM SIGKDD explorations newsletter*, *2*(2), 34-38.