DATA CLUSTERING TECHNIQUE IN BIG DATA AND MINING - A PERSPECTIVE STUDY

*****Pradip Kumar Yadava and **Dr. Rajeev Kumar Yadav
[1]Research Scholar, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
[2]Professor, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Corresponding author Email: - pradip.yadava@gmail.com

Abstract: The need to understand large, complex, information rich data sets is common to all fields of studies in this current information age. Given this tremendous amount of data, efficient and effective tools need to be present to analyze and reveal valuable knowledge that is hidden within the data. Clustering analysis is one of the popular approaches in data mining and has been widely used in big data analysis. The goal of clustering involves the task of dividing data points into homogeneous groups such that the data points in the same group are as similar as possible and data points in different groups are as dissimilar as possible. The importance of clustering is documented in pattern recognition, machine learning, image analysis, information retrieval, etc. Due to the difficulties of parallelization of the clustering algorithms and the inefficiency at large scales, challenges for applying clustering techniques in big data has arisen. The question is how to deploy clustering algorithms for this tremendous amount of data to get the clustering result within a reasonable time. This chapter provides an overview of the mainstream clustering techniques proposed over the past decade and the trend and progress of clustering algorithms applied in big data. Moreover, the improvement of clustering algorithms in big data are introduced and analyzed. The possible future for more advanced clustering techniques are illuminated based on today's information era. Cluster Analysis in Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups. In the process of clustering in data analytics, the sets of data are divided into groups or classes based on data similarity. Then each of these classes is labelled according to their data types. Going through clustering in data mining example can help you understand the analysis more extensively. In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. After the classification of data into various groups, a label is assigned to the group. It helps in adapting to the changes by doing the classification.So if we were to define clustering in data mining, then we can say that the process of cluster in data mining is basically comprising a set of abstract objects into groups of similar objects. The process of dividing and storing them in these groups is known as cluster analysis.

[Yadava, P.K. and Yadav, R.K. **DATA CLUSTERING TECHNIQUE IN BIG DATA AND MINING - A PERSPECTIVE STUDY**. *The International Journal of Interpretation, Observation and Analysis*, 2024; Volume 2, Issue 1:18-23. (April-June). ISSN 2349-0713, Peer-reviewed (online/offline), Refereed, Indexed and International Journal (Since 2013), Global Impact Factor: 5.776]

Keywords: Women writer, Indian Writers, Feminism in Literature

Introduction: An overwhelming flow of data in structured, unstructured or heterogeneous format has been accumulated due to the continuous increase in the volume and detail of data captured by organizations, such as social media, government, industry and science. These massive quantities of data are produced because of the growth of the web, the rise of social media, the use of mobile, and the information of Internet of Things (IoT) by and about people, things, and their interactions. The big data era has arrived. Big data becomes the most influential force in daily life. According to the IDC reports, the digital universe is doubling in size every two years and it will reach 44 zettabytes by 2020 [1]. How to store huge amounts of data is not the biggest problem anymore. But how to design solutions to understand this big amount of data is a major challenge. Operations such as analytical operations, process operations, retrieval operations, are very difficult and hugely time consuming because of this massive volume of data. One solution to overcome these problems is the use of data mining techniques in discovering knowledge from big data. Data mining [2] is called exploratory data analysis, among other things. It is an analytic process designed to explore data. Data mining aims to search for consistent patterns or systematic relationships between variables. It then validates the findings by applying the detected patterns to new subsets of data. Although the hidden patterns are derived from heterogeneous data in big data mining, these hidden patterns can still be reviewed as structured knowledge. The structured knowledge is combined with human knowledge of decision makers that are heterogeneous or

unstructured and upgraded into intelligent knowledge [3].

Clustering is a kind of unsupervised machine learning technology, which is used to mine the intrinsic similarity of data and divide the data set into several subsets. Each data subset is a cluster, the samples within the cluster are similar to each other, and the samples between different clusters are not similar. In general, the similarity of samples is characterized by Euclidean distance, Markov distance, Manhattan distance, Pearson distance, Chebyshev distance, cosine similarity, Jaccard similarity and probability density. Clustering techniques have been widely used in real life, such as customer grouping in commercial activities, gene sequence classification in bioinformatics, spam identification in the Internet, and analysis of industry electricity usage behavior in the electricity market. With the advent of the era of big data, data collection and storage has become relatively easy and convenient. Large-scale data sets of GB-level and even TB-level storage are emerging one after another. The size of data sets of big data is growing at an unimaginable speed, which brings great challenges to data processing. Therefore, clustering research for large data sets is constantly emerging. So far, clustering algorithms for different types of small and medium-sized data sets have made a historic breakthrough in clustering accuracy. However, these algorithms still have many problems when dealing with large data sets. The main defects are high computational complexity and long computing time, which is unacceptable.

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

Simply it is the partitioning of similar objects which are applied to unlabelled data.

General Information of Clustering Analysis

Clustering is one of the most fundamental tasks in exploratory data analysis that groups similar data points in an unsupervised process. Clustering techniques have been exploited in many fields including in many areas, such as data mining, pattern recognition, machine learning, biochemistry and bioinformatics [9]. The main process of clustering algorithms is to divide a set of unlabeled data objects into different groups. The cluster membership measure is based on a similarity measure. In order to obtain a high quality partition, the similarity measure between the data objects in the same group is to be maximized, and the similarity measure between the data objects from different groups is to be minimized. Most of the clustering task uses an iterative process to find locally or globally optimal solutions from a high-dimensional data sets. In addition, there is no unique clustering solution for real-life data and it is also hard to interpret the ‘cluster’ representations [9]. Therefore, the clustering task requires much experimentation with different algorithms or with different features of the same data set. Hence, how to save 6 computational complexity is a significant issue for the clustering algorithms. Moreover, clustering very large data sets that contain large numbers of records with high dimensions is considered a very important issue nowadays. Most conventional clustering algorithms suffer from the problem that they do not scale with larger sizes of data sets, and most of them are computationally expensive with regards to memory space and time complexities. For these reasons, the parallelization of clustering algorithms is a solution to overcome the aforementioned problems, and the parallel implementation of clustering algorithms is inevitable.

More importantly, clustering analysis is distinguished from other analysis [9]. Clustering analysis belongs to the so called unsupervised learning category. The main goal is to divide a set of unlabeled data sets into several groups based on the conceptual or hidden properties of the input data sets. In other words, clustering analysis is unsupervised 'nonpredictive' learning. It divides the data sets into several clusters based on a subjective measurement. Clustering analysis is unlike supervised learning and it is not based on a 'trained characterization'. In general, there is a set of desirable features for a clustering algorithm [9,10]: scalability, the temporal and spatial complexity of the algorithm should not explode on large data sets. Robustness, the outliers in the data set should be detected during the process. Order insensitivity, the ordering of the input data should not affect the outcome of the algorithm. Minimum user-specified input, the number of user-specified parameters should be minimized. Arbitrary-shaped clusters, the clusters can be shaped arbitrarily. Point proportion admissibility, different clustering algorithms produce different results with different features. Hence, a clustering algorithm should be chosen such that duplicating the data set and the re-clustering task should not change the clustering results.

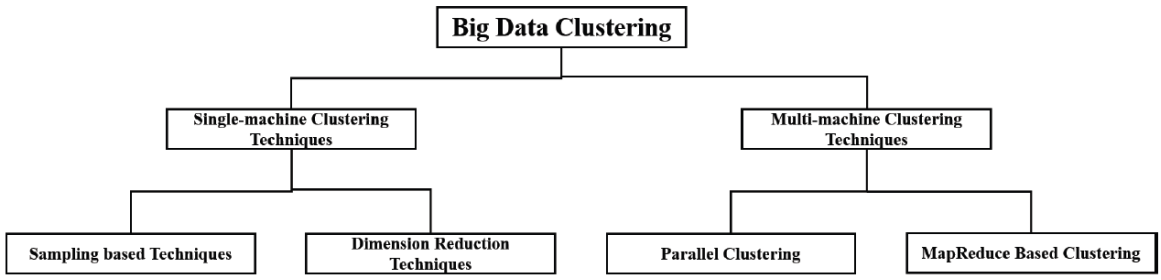

Figure 1: A list of big data clustering techniques.

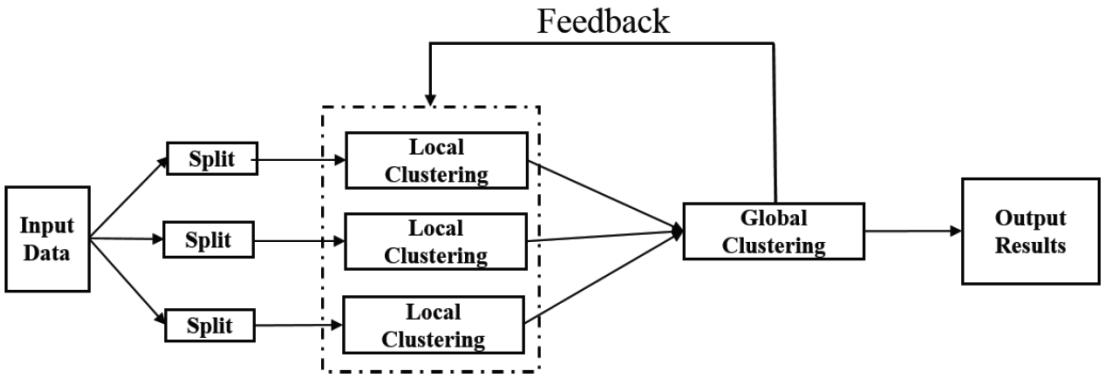

Figure 2: The general framework of most parallel and MapReduce clustering.

Big data clustering analysis

The speed of information growth exceeds the Moore's Law at the beginning of this new century. Given this tremendous amount of data, efficient and effective tools need to be present to analyze and reveal valuable knowledge that is hidden within the data. Clustering is one of the popular approaches in data mining and has been widely used in big data analysis. The goal of clustering involves the task of dividing data points into homogeneous groups such that the data points in the same group are as similar as possible and data points in different groups are as dissimilar as possible. However, conventional clustering techniques cannot cope with this huge amount of data because of their high complexity and computational cost [8]. The question for big data

clustering is how to scale up and speed up clustering algorithms with minimum sacrifice to the clustering quality. Therefore, an efficient processing model with a reasonable computational cost of this huge, complex, dynamic and heterogeneous data is needed in order to exploit this huge amount of data. Single machine clustering techniques and multiple machine clustering techniques are two most popular big data clustering techniques. Single-machine clustering algorithms run in one machine and can use resources of just one single machine while the multi-machine clustering [8] techniques can run in several machines with access to more resources. Multi-machine clustering techniques have become more popular due to the better scalability and faster user response time.

Properties of Clustering :

1. Clustering Scalability: Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can't get the appropriate result which would lead to wrong results.

2. High Dimensionality: The algorithm should be able to handle high dimensional space along with the data of small size.

3. Algorithm Usability with multiple data kinds: Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. Dealing with unstructured data: There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. Interpretability: The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Methods:

The clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method: It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

- **Agglomerative Approach:** The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.
- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.
- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After

grouping data objects into microclusters, macro clustering is performed on the microcluster.

Density-Based Method: The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

Grid-Based Method: In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Model-Based Method: In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

Constraint-Based Method: The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Key Technologies for clustering in big data

Unlike traditional clustering algorithms, volume of data must be taken into account when clustering big data because this requires substantial changes in the architecture of storage system. In addition, most of the traditional clustering algorithms are designed to handle either numeric or categorical data with limited size. On the other hand, big data clustering deals with different types of data such as image, video, sensors, mobile devices, text etc. [11, 12]. Moreover, the velocity of big data requires the big data clustering techniques to have a high demand for online processing of data. At the high level of the five category clustering algorithms, most of the algorithms have a similar procedure. These algorithms start with some random initialization and follow some iterative process until some convergence criteria are met. For example, partitioned clustering such as the k-means algorithm, starts with randomly chosen k centroids and reassigns each data point to the closest cluster centroids in an iterative process. Thus, the issue of big data clustering is how to speed up and scale up the clustering algorithms with the minimum sacrifice to the clustering quality. There are three ways to speed up and scale up big data clustering algorithms. 9 The first way is to reduce the iterative process using sampling-based algorithms. Samplingbased algorithms perform the clustering based on a sample of the datasets instead of using on the whole dataset. Complexity and memory space needed for the process decreases in samplingbased algorithms because computation only needs to take place for smaller sample dataset. PAM, CLARA and CLARANS [9-12] are proposed to fight with the exponential search space in the Kmedoid clustering problem. The second way is to reduce the data dimension using randomized techniques. Dimensionality of the dataset is another aspect which influences the complexity and speed of clustering algorithms. Random projection and global projection are used to project dataset from a high dimensional space to a lower dimensional space [8, 12]. CX/CUR, CMD and Colibri [8, 12] are dimension reduction techniques which are proposed to reduce long execution times of big data clustering. The last way is to apply parallel and distributed algorithms use multiple machines to speed up the computation in order to increase the scalability. Parallel processing applications include conventional parallel applications and data-intensive applications. The conventional parallel applications assume that data can be fit into the memory of distributed machines. Data intensive applications are I/O bound and devote the largest fraction of execution time to the movement of data. OpenMP, MPI [13], and MapReduce are common parallel processing models for computing data-intensive applications.

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.
2. It can be used for exploratory data analysis and can help with feature selection.
3. It can be used to reduce the dimensionality of the data.
4. It can be used for anomaly detection and outlier identification.
5. It can be used for market segmentation and customer profiling.

Conclusion After continuous research, some research results have been achieved in big data, such as big data search, big data storage, big data mining, etc., but still cannot meet the needs of current big data. Researching real-time, highly robust new high-efficiency clustering algorithms for big data has become a key task to be solved in the deep exploration of the hidden value of big data. In the field of data mining, the final result of many clustering algorithms is sensitive to the correct setting of parameters, which leads to these algorithms far from being called mature and practical intelligent machine learning algorithms. In the big data environment, it is necessary to study and design a more efficient intelligent automatic clustering algorithm. Therefore, the clustering algorithm for big data needs constant research to meet the needs of current big data.

References:

[1] Executive Summary Data Growth, Business Opportunities, and the IT Imperatives An ICD report. Retrieved from www.emc.com/leadership/digital-universe/2014iview/executivesummary.htm.

[2] H. A. Edelstein. Introduction to data mining and knowledge discovery (3rd ed). Potomac, 25 MD: Two Crows Corp. 1999.

[3] Z. Xu and Y. Shi, Exploring Big Data Analysis: Fundamental Scientific Problems. Annals of Data Science, 2(4), 363-372, 2015.

[4] C. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347, 2014.

[5] D. Laney: 3D Data Management Controlling-Data Volume, Velocity and Variety (February 2001).

[6] Y. Demchenko, P. Grosso, C. De Laat and P. Membrey, Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 IEEE International Conference on (pp. 48-55).

[7] A. Fahad, N. Alshatri, , Z.Tari, A. Alamri, I.Khalil, A. Y. Zomaya, and A. Bouras, A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279, 2014.

[8] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, Big data clustering: a review. In International Conference on Computational Science and Its Applications (pp. 707- 720). Springer International Publishing, 2014.

[9] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons, 2009.

[10] W. Kim, Parallel clustering algorithms: survey. Parallel Algorithms, Spring, 2009.

[11] C. C.Aggarwal, C. K. Reddy, Data clustering: algorithms and applications. CRC Press, 2013.

[12] R. T. Ng, J. Han, CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016, 2002.

[13] J. A. Zhang, Parallel Clustering Algorithm with MPI–Kmeans. Journal of computers 8.1 (2013): 10-17.

[14] I. Jolliffe, Principal component analysis. John Wiley & Sons, Ltd, 2002.

[15] J. Yadav, D. Kumar, Subspace Clustering using CLIQUE: An Exploratory Study.

[16] Q. Gu, J. Zhou, Co-clustering on manifolds. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 359-368), 2009.

[17] S. Kantabutra and A. L. Couch, Parallel K-means clustering algorithm on NOWs. NECTEC Technical journal, 1(6), 243-247, 2000.