International Journal on Applications in Engineering and Technology
Volume 3: Issue3 : December 2017, pp 1 7 - 21 www.aetsjournal.com *ISSN (Online) : 2455 - 0523*

A Cross-Sectional Survey On Parallel Data Partitioning Clustering Algorithms[1]

G.Karthikeyan , Dr. Komarasamy.G

Abstract— Data analysis is a huge and immense process used as a universal method in modern science research, such as information communication, algorithm oriented computer science and human interaction biological information science. The concept of Clustering, as the basic data analysis, plays a significant role. On one hand, many tools for cluster analysis in mining have been created, along with the huge volume of information increase and subject label intersection of data. On the other hand, each clustering algorithm has its own strengths and weaknesses, due to the complexity of handling information and large volume of data produced from machines. In this review paper, we begin at the definition of clustering, take the basic elements involved in the clustering process, classification and analyze the clustering algorithms from two perspectives, the traditional ones and the modern ones.

Keywords -- K-means, K-medoids, BIRCH, DBSCAN, OPTICS, STING, and WaveCluster

I. INTRODUCTION

Clustering is a data mining technique that groups data into meaningful subclasses, known as clusters, such that it minimizes the intra-differences and maximizes inter-differences of these subclasses. Well-known algorithms include K-means, K-medoids, BIRCH, DBSCAN, OPTICS, STING, and WaveCluster. These algorithms have been used in various scientific areas such as satellite image segmentation, noise filtering and outlier detection, unsupervised document clustering, and clustering of bioinformatics data. Existing data clustering algorithms have been roughly categorized into four classes: partitioning-based, hierarchy-based, grid-based, and density-based. We provide parallel implementations for three clustering algorithms, OPTICS, DBSCAN, and single-linkage hierarchical agglomerative clustering.

Mr.G.Karthikeyan, Assistant Professor / CSE, JKK Munirajah College Of Technology(Email Id : gkarthikeyancse@gmail.com,)

Dr. Komarasamy.G , Assistant Professor (senior Grade) , Department of CSE, Bannari Amman Institute of Technology, India. (Email Id: gkomarasamy@gmail.com)

1) RANKING MODEL

Learning based information recovery method Letor), learning to Rank focused in learning a ranking model with some labeled documents with their significance to some uncertainty, whe0re the model is optimistically capable of ranking the documents returned to an unsystematic new query automatically. There are some various machine learning methods, e.g., Ranking SVM Rank Boost, Rank Net, List Net, Lambda Rank, based on this, the learning to rank algorithms have already exposed their promising performances in information retrieval, particularly web search.

We know that modem database environments are very large, and the number of topically relevant documents to any one request may easily exceed the number of documents a user is interested to scrutinize.

On the other hand, as the domain-specific search Engines coming out with more attentions have moved from the broad-based search to specific verticals, for tracking information constraint to a certain domain.

2) CLUSTERING

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. Clustering is defined as dividing input data sets called “clusters”.

As unsupervised learning tasks, clustering tasks have been exploited in many fields including machine learning, data mining, video processing, biochemistry and bioinformatics.

Depending on the data properties or the purpose of clustering, different types of clustering algorithms have been developed, such as, partitioned, hierarchical, graph-based clustering etc. There are no predefined class label exists for the data points.

3) CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes.

The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.

4) DISTANCE AND SIMILARITY

Distance (dissimilarity) and similarity are the basis for constructing clustering algorithms. As for quantitative data features, distance is preferred to recognize the relationship among data. And similarity is preferred when dealing with qualitative data features.

5) EVALUATION INDICATOR

The main purpose of evaluation indicator is to test the validity of algorithm. Evaluation indicators can be divided into two categories, the internal evaluation indicators and the external evaluation indicators, in terms of the test data whether in the process of constructing the clustering algorithm.

The internal evaluation takes the internal data to test the validity of algorithm. It, however, can't absolutely judge which algorithm is better when the scores of two algorithms are not equal based on the internal evaluation indicators.

6) CLUSTERING IN DATA MINING

Clustering is a technique in which a given data set is divided into groups called clusters in such a manner that the data points that are similar lie together in one cluster.

Clustering plays an important role in the field of data mining due to the large amount of data sets.

This paper reviews the various clustering algorithms available for data mining and provides a comparative analysis of the various clustering algorithms like DBSCAN, CLARA, CURE, CLARANS, K-Means etc.

II. VARIOUS CLUSTERING ALGORITHMS:

1. Partitioning Clustering Algorithms
2. Hierarchical Clustering Algorithms
3. Evolutionary Clustering Algorithms
4. Density-based Clustering Algorithms
5. Model-based Clustering Algorithms
6. Graph-based Clustering Algorithms

TRADITIONAL CLUSTERING ALGORITHMS

The traditional clustering algorithms can be divided into 9 categories which mainly contain 26 commonly used ones, summarized in Table

Category	Typical algorithm
Clustering algorithm based on partition	K-means, K-medoids, PAM, CLARA, CLARANS
Clustering algorithm based on hierarchy	BIRCH, CURE, ROCK, Chameleon
Clustering algorithm based on fuzzy theory	FCM, FCS, MM
Clustering algorithm based on distribution	DBCLASD, GMM
Clustering algorithm based on density	DBSCAN, OPTICS, Mean-shift
Clustering algorithm based on graph theory	CLICK, MST
Clustering algorithm based on grid	STING, CLIQUE
Clustering algorithm based on fractal theory	FC
Clustering algorithm based on model	COBWEB, GMM, SOM, ART

1) CLUSTERING ALGORITHM BASED ON PARTITION

The basic idea of this kind of clustering algorithms is to regard the center of data points as the center of the corresponding cluster. K-means and K-medoids are the two most famous ones of these kinds of clustering algorithms.

The core idea of K-means is to update the center of cluster which is represented by the center of data points, by iterative computation and the iterative process will be continued until some criteria for convergence is met. K-mediods is an improvement of K-means to deal with discrete data, which takes the data point, most near the center of data points, as the representative of the corresponding cluster.

2) CLUSTERING ALGORITHM BASED ON HIERARCHY

The basic idea of this kind of clustering algorithms is to construct the hierarchical relationship among data in order to cluster. Suppose that each data point stands for an individual cluster in the beginning, and then, the most neighboring two clusters are merged into a new cluster until there is only one cluster left.

Typical algorithms of this kind of clustering include BIRCH, CURE, ROCK, and Chameleon. BIRCH realizes the clustering result by constructing the feature tree of clustering, CF tree, of which one node stands for a sub cluster. CF tree will dynamically grow when a new data point comes. ROCK is an improvement of CURE for dealing with data of enumeration type, which takes the effect on the similarity from the data around the cluster into consideration. Chameleon, at first, divides the original data into clusters with smaller size based on the nearest neighbor graph, and then the clusters with small size are merged into a cluster with bigger size, based on agglomerative algorithm, until satisfied.

3) CLUSTERING ALGORITHM BASED ON FUZZY THEORY

The basic idea of this kind of clustering algorithms is that the discrete value of belonging label, {0, 1}, is changed into the continuous interval [0, 1], in order to describe the belonging relationship among objects more reasonably.

4) CLUSTERING ALGORITHM BASED ON DISTRIBUTION

The basic idea is that the data, generated from the same distribution, belongs to the same cluster if there exists several distributions in the original data. The typical algorithms are DBCLASD and GMM.The core idea of DBCLASD, a dynamic incremental algorithm, is that if the distance between a cluster and its nearest data point satisfies the distribution of expected distance which is generated from the existing data points of that cluster, the nearest data point should belong to this cluster.

The core idea of GMM is that GMM consists of several Gaussian distributions from which the original data is generated and the data, obeying the same independent Gaussian distribution, is considered to belong to the same cluster

5) MODERN CLUSTERING ALGORITHMS

The modern clustering algorithms can be divided into 10 categories.

Category	Typical algorithm
Clustering algorithm based on kernel	kernel K-means, kernel SOM, kernel FCM, SVC, MMC, MKC
Clustering algorithm based on ensemble	CSPA, HGPA, MCLA, VM, HCE, LAC, WPCK, sCSPA, sMCLA, sHBGPA
Clustering algorithm based on swarm intelligence	ACO_based(LF), PSO_based, SFLA_based, ABC_based
Clustering algorithm based on quantum theory	QC, DQC
Clustering algorithm based on spectral graph theory	SM, NJW
Clustering algorithm based on affinity propagation	AP
Clustering algorithm based on density and distance	DD
Clustering algorithm for spatial data	DBSCAN, STING, Wavecluster, CLARANS
Clustering algorithm for data stream	STREAM, CluStream, HPStream, DenStream
Clustering algorithm for large-scale data	K-means, BIRCH, CLARA, CURE, DBSCAN, DENCLUE, Wavecluster, FC

6)PARALLEL CLUSTERING

Clustering is grouping input data sets into subsets, called 'clusters' within which the elements are somewhat similar. In general, clustering is an unsupervised learning task as very little or no prior knowledge is given except the input data sets. The tasks have been used in many fields and therefore various clustering algorithms have been developed.

Clustering task is, however, computationally expensive as many of the algorithms require iterative or recursive procedures and most of real-life data is high dimensional. Therefore, the parallelization of clustering algorithms is inevitable, and various parallel clustering algorithms have been implemented and applied to many applications.

7) CLUSTERING ANALYSIS COMPONENTS

A clustering task needs several essential steps as the followings:

1) Pattern representation,
2) Measurements appropriate to the data domain,
3) Clustering or grouping,
4) Data abstraction and
5) Assessment of output.

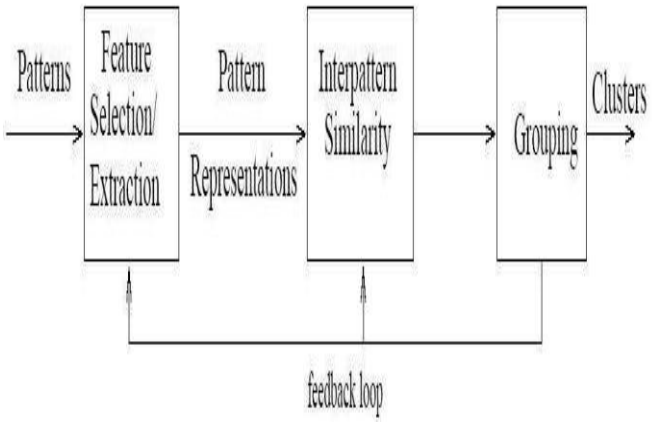

Figure 1: Clustering procedure with a feedback pathway

8) CLUSTERING TASK

The procedure of clustering task similarly as the followings:

1) Feature selection or extraction
2) Clustering algorithm design or selection
3) Cluster validation
4) Results interpretation.

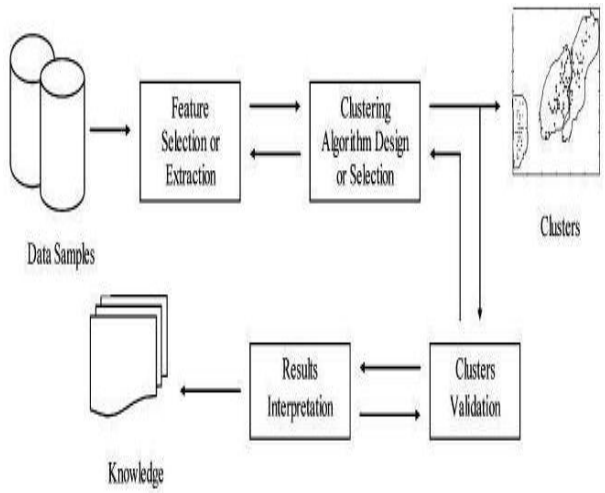

Figure 1: Clustering Task

III. PARALLEL IMPLEMENTATION OF CLUSTERING ALGORITHMS

1. Parallel Partitioning Clustering Algorithms
2. Parallel Hierarchical Clustering Algorithms
3. Parallel Evolutionary Clustering Algorithms
4. Parallel Density-based Clustering Algorithms
5. Parallel Model-based Clustering Algorithms
6. Parallel Graph-based Clustering Algorithms

APPLYING CLUSTERING ALGORITHMS FOR PARALLELIZATION

Clustering algorithms and parallelization is closed related because they are useful applying the 'divide and conquer' strategy in algorithms so that they reduce the time complexities. Therefore, while some parallel techniques used for efficient clustering algorithms, some clustering algorithms used for parallel tasks as well. In this section, we provide a number of literatures which used clustering algorithms for efficiency.

IV. CONCLUSION

In this study, we reviewed the papers about clustering algorithms and the corresponding parallel clustering algorithms. Clustering algorithms are categorized to five classifications: partitioning, hierarchical, evolutionary approach, dense-based, model-based and graph-based algorithms. Partitioning and hierarchical algorithms are the most popular clustering algorithms and many of parallel version of the algorithms also have been developed.

Parallelism in the clustering algorithms has been used both for efficient clustering strategy and for the efficient distance-computation.

K-means algorithm is the representative partitioning clustering algorithm and the parallelization was implemented mostly in message-passing model. For efficient data communication, some of the algorithms used the properties of specific interconnection network topology including hypercube or butterfly while others use master-slave configuration to parallelize the distance computation or the cluster assignment.

Hierarchical clustering algorithms are more expensive than others in computation as it involves computing the level of clustering as well. Parallel version of those

algorithms with single-linkage metric use hypercube network to reduce the computation of MST, or use message-passing architecture to efficiently compute the similarities and to update the memberships of each data.

V. CLUSTERING METHODS

The goal of clustering methods is to group elements sharing the same information. The concept of similarity represents the essence of clustering. Data are represented by computers in binary format by using two distinct symbols: 1 and 0. Therefore, two files are rarely identical, but could be similar to a certain degree, which means that the real problem of clustering remains to gather the elements which are most similar between them but less similar to all the others. This observation raises an important question if it is always possible to classify data. The answer is positive, but in certain cases the clustering might be not relevant. The clustering methods can be divided into the following three categories:

– Distance methods
– Characters methods
– Quadruplets methods

In order to achieve the objective of this work, only the methods of distance and the methods of quadruplets (which also use criteria based on the notion of distance) were considered for analysis and implementation, for the reason that the normalized compression distance (NCD) can only be used with methods based on distance.

REFERENCES

[1] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2015.
[2] S. Kantabutra and A.L. Couch. Parallel K-Means Clustering Algorithm on NOWs. *NECTEC Technical Journal*, 1(1), 2009.
[3] Fazilah Othman, Rosni Abdullah, Nur'Aini Abdul Rashid, and Rosalina Abdul Salam. Parallel K-Means Clustering Algorithm on DNA Dataset. In *PDCAT*, pages 248-251, 2014.
[4] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Partitioning-based clustering for web document categorization," Decision Support Systems 27:329- 341, 1999
[5] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Partitioning-based clustering for web document categorization," Decision Support Systems 27:329-341, 1999.