International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195270

A Systematic Algorithm for Data Cluster Using Map-Reduce Approach

Kechika. S[1], Sapthika. B[1], Keerthana. B[1], Abinaya. S[1], Abdulfaiz. A[2]
[1]Scholar, Department of CSA, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India
[2]Professor, Department of CSA, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT

We have been studying the problem clustering data objects as we have implemented a new algorithm called algorithm of clustering data using map reduce approach. In cluster, main part is feature selection which involves in recognition of set of features of a subset, since feature selection is considered as a important process. They also produces the approximate and according requests with the original set of features used in this type of approach. The main concept beyond this paper is to give the outcome of the clustering features. This paper which also gives the knowledge about cluster and it's own process. To processing of large datasets the nature of clustering where some more concepts are more helpful and important in a clustering process. In a clustering methodology where more concepts are very useful. The feature selection algorithm which affects, the entire process of clustering is the map-reduce concept. since, feature selection or extraction which is also used in map-reduce approach. The most desirable component is time complexity where efficiency concerns in this criterion. Here time required to find the effective features, where features of quality subsets is equal to effectiveness. The complexity to find based on this criteria based map-reduce features selection approach, which is proposed and evaluated in this paper.

Keywords : Features Selection, Features Clustering, Map-Reduce, Map-Reduce Approach

I. INTRODUCTION

Generally data mining is the process of partitioning a set of data(or objects) into a set of meaningful sub-classes, called clusters. Cluster is similar to classification in that data are grouped. Unlike classification, the groups are non-predefined. Instead of grouping is accomplished by finding similarities between data according to the characteristics in the actual data. The groups are called clusters. Different groups are equal from one another. Commonly used functional programming is inspired by the map-reduce functions and consists of two stages only one single data is used in contrast to many conventional clustering based on algorithm, to produce featured outputs. Whereas we introduce a map-reduce approach for clustering. In business, marketing purpose, cluster analysis is used to discover, implement and characterize customer segments.

Four types of clustering algorithms,

1. Programs occurring when clustering is applied to real world database.
2. Outlier handling is difficult.
3. Dynamic data in the database implies the cluster membership may change overtime.
4. Interpreting the semantic meaning of each clusters may difficult. The exact meaning is of each cluster is may not clear.

5. Another related issue is which type of data should be used for clustering. A prior knowledge to unsupervised learning.
6. The important issue in clustering is that, how to determine similarity between two objects, so that within cluster. They can be formed with low and high similarity between objects.
7. Generally, to measure similarity and dissimilarity between objects, a distance such as, Manhattan Minkowski Euclidean are used.
8. The distance function returns the lower bounded value which is similar to one another distance measured in some kind of technique used in data mining.
9. A data analysis which includes data mining, image analysis and that is defined in data clustering.

There are many types of clustering algorithms. The actual concept in hierarchical algorithms is they actually creates set of clusters set of algorithms different in how the sets are created. A dendrogram which is of tree data structure can be used to illustrate the hierarchical clustering technique and the sets of many types of clusters. The root in the dendrogram tree consists of one cluster. The leaves in the dendrogram consists of a single element closer. Internal nodes in a dendrogram represent new clusters formed by merging the clusters. Each level in the tree is associated with the distance measure that was used to merge clusters hierarchical may be agglomerative in nature that follows bottom-up approach, which means they built clusters by consecutive by merging the smaller ones. It can be hostility in nature, followed by top-down approach. In case of space and time complexity, clustering of data can more expensive. By repeating this clustering of data more experienced might be acquired. In terms, increased quality more computation speed of distributing parallelizing the data becomes more attractive often.

Fig.1.What is Data Mining& Classification

A framework is introduced map-reduce framework, it is useful in resolving various kinds of distribution problem. Map-reduce framework consists of mapping and reduces functionalities, it can be made in two steps. It has two simplest step process. In the first step of framework which can be divided into several types of identical and independent parts that can be specific to map tasks. The output can be in the form of key-value pairs on the second step of map-reduce method. The reduced parts can have the results from the map task and certain pair-key is processed. Thus the power of framework comes from certain fact, the map-reduce framework, it is adjective to the distributed sorting platform.

Definition

The total and main goal of the data mining process is to obtain information from the hidden data set and convert it into a understandable structure for future use. Clustering is similar to classification in that data are group in the same way (or) the data are similar.

Hierarchical algorithm find future successive clusters by previously obtained clusters. It can be either be agglomerative or divisive. Since agglomerative algorithm begin with objects as singleton clusters and also merge with other clusters to create final clusters. Divisive algorithm can begin with entire data set of

objects and divided into smaller clusters. Large data sets can be computed using many computes in a map-reduce framework.

a) Pattern representative:

It represents (or) denotes, number of available patterns, the number of classes, the number, types and some of the features needed clustering algorithm. Its characteristics can be sub-divided into many types, The pattern presents the knowledge that can be well resolved by the humans. Which are valid with some grades of sustainability, impressively useful, which has some validity about the useful was unusual. Measure of pattern can be applied performance of the discovery procedure. According to the Varieties of database mined, data mining can be separated. Large database which are effective and efficient provides many requirements and challenges to researchers and scholars.

The main issues involved in data mining methodology user dealing presentation, and ascendable and the processing of large characters.

A data mining system can have ability to get more than thousands of pattern and its rules. They also contain same parallelization, ability and ascendable algorithms. These algorithms are used to obtain information from the large and big amount of data in a database.

The large amount of multiple database, and some of their data mining methods are some factors which encourages the implementation and development of paralleled and distributed data mining algorithms. Thus the algorithm clusters data into divisions, they are in parallel. The results are then combined.

II. METHODS AND MATERIAL

A. Feature Selection/Feature Extraction

Feature selection is the process of finding the most effectual subset of exact feature to use in clustering. It is a function of one or more changes of one of one input to produce new major features. Both of these techniques can be implied to keep hold of appropriate set of features in the building process. The main assumption in the process of taking out a subset of relevant characteristics for function in model building. The currently selected future which provides information, future selection partitionized into Four types: wrapping, hybrid, filter, and embedded method. Where wrapper method is used for optimization of the looping process.

B. Measures of Similarities

Clustering algorithm is the process of finding differences of similarity between two objects. For this, the distance function is used where these function takes two objects as input and returns real number as positive corresponding distance between objects is smaller. Depending upon the specific application several popular functions are available and one should be needed to chosen for clustering problem. Based on the different types of attributes length function is chosen used to an instance in clustering.

MAP Reduce Approach:

Map reduce step can solve a complex problem. Every steps takes an output from a previous step in map reduce.

Data preperation:

The collecting of huge amount of file each containing a data of one set is called data set. Record identification number of the record of these file should be in the first line.

Map step:

The average output is recorded ID as key and returned as value. Each mapper is maintained a collection of canopy enter cadidates it has learned far. The mapper determines if each successive records

with in the distance threshold of any already determined canopy center candidate. The output which is intermediate is send to the reducer has record ID as the list of retired-rating pairs as the value.

Reduce step:

The outcome of the reduced step will simply output record ID as the key concentrate the rater ID's for recording comma separated list. The mapper's repeat same steps to reduce. It takes out those which are inside the same threshold limit, as it meets the canopy center ID's. Its also removes duplication.

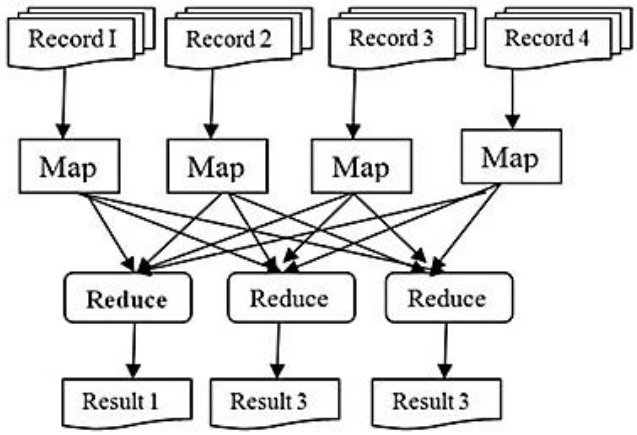

Fig.2.Map-Reduce Functionality

III. RESULTS AND DISCUSSION

A) ALGORITHM IMPLEMENTATION:

As we have implemented a algorithm maRc, which is known map reduce algorithm to clustering data. An effective module which is introduced in a beginning step is known as map, where as another module which is introduced can use in merging the in between data which results in map phase, the in between data that can be processed which can give efficient and effective features. During prediction, it processed the clustered data With help of key and value pairs and during reduction the in between data can be processed with key-value collection and major function can be performed.

Algorithm1:

MaRc algorithm with map reduce

Map function {(key value) /*(Record_id, Record_value),Tα*/}

Input: cluster data (key, value)

Ouput: in between data sets Tb¥B

1:map((cons key & key) Tα)/*(Record_id,Record_value)*/

//mapping (split) key pairs

{

2: For each{ key, value, key} in {(key, value) Tb}

3: Pα = f(key,Tα)

4: For each key€key^

5: Emit(key^,Pb) in the in between data

6: Tb = Tb U (key, value, Pb)

7: }

Reduce function:

{(key, key^,value)/*(Record_id,Record_value),Tα,*/}

Input: An in between data sets Tb.

Output: Estimated coefficient Eα.

1. Reduce {(key^, value) ,Pb} // in between data
2. {
3: Eα = {(key^,value) Pb}
4: for each (key^,value)€Eα}
5: Eα = Tb n (key^,Pb) // estimated co-efficient
6: }

Algorithm : MARC algorithm

Where (key, value) where initial data cluster pairs.

Tb is the total data sets as input.

Pb is the intermediate data (key^,value).

Eα is the estimated co-efficient [outputs].

Key^ is the key pair of intermediate (or) inbetween data to be reduce.

How many maps and reduces?

A fact that reduce mapreduce framework is to determine how many maps and scales down to use for optimal operation it is useful if one requires

outfiles for to be used in as input in another map reduce, also ensures correct result in some other calculations. At the same time, canopy selection map reduce where number of reduces to be set in one order for it to function right. A machine which will not respond is considered as 'dead' both amp and reduce machines needs to be re-executed and becomes eligible for scheduling.

The execution of many map tasks. Which is the power of map reduce framework n parallel on a data sets and the outputs as in between key-value pairs. Each reduce phase only receives as well as processes data for a particular key at an instance and outputs information it process as it key-values. Hence, distributed sorting framework is most basic usage in a map reduces. It becomes attractive as it permits a coder to write software for implementation on a computing cluster with help of knowledge of parallel (or) distributed computing usage of more reduce tasks reduce lower the cost failures, out increases the overhead of framework load (balancing).

a) Joins:

A join which is a popular operator that is not so well deact with my map and reduce function. So, far map reducing is designated for processing a one single input, the supports of joining that requires more than one inputs where as map reduce has been an open issues. Map reduce has divided with joins into two groups.

Map-side join and reduce-side join.

b) map-side join:

A map-side join is map-merge join that works similar to sort-merge join. The beginning, two inputs in the first two reactional inputs are partitioned and sorted on the join keys. Broadcast join method is the another map-side join method, when it is applicable size of a reaction is small. The smaller reaction is broadcast to each mapper and kept into memory.

c) reduce-side join:

Reduc-side join is a repartitioned join. Each mapper tags is to identify which related the rows comes from. During shuffling, rows of the tags, which have same key values are copied.

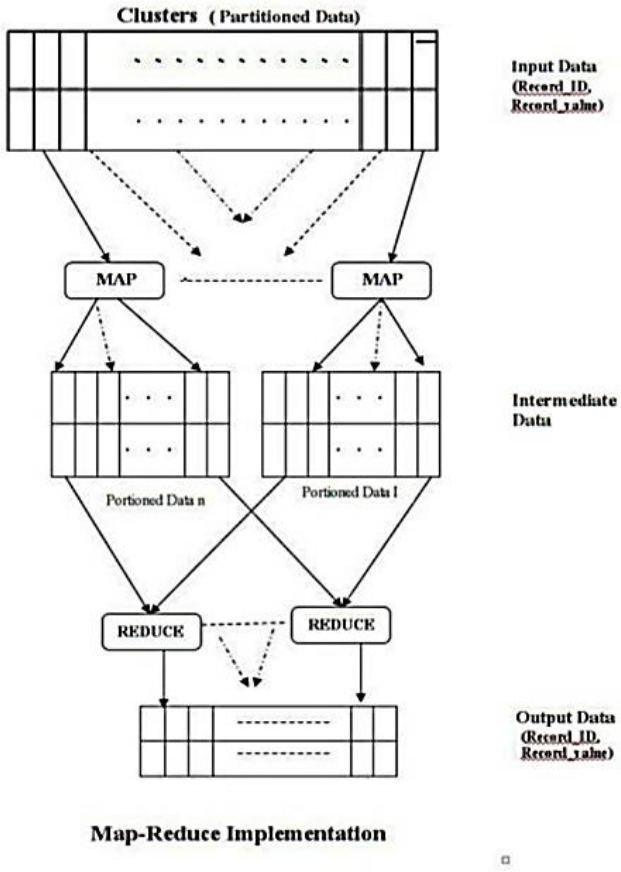

Fig.3. Map-Reduce Implementation

APPLICATION

In computing aggregate map reduce is simple and effective. In dbms, it is a great deal Comparing with "filtering them group bu aggregation " information processing is the major advantage of map reduce framework. The map reduce model is simple but also expensive. By without specify physical distribution of the job across nodes. A program defines his job with only map and reduce functions. Data model and schema does not have any dependency on flexible map reduce.

FRAMEWORK

Joining and reducing actually which running inside the same reduce framework. The map reduce join with two consecutive mr jobs is also proposed to avoid modifying the map reduce framework. In a multi-way join, join-chain represented as leaf-deep tree. The previous joiners transfers transfers joiner row to the next joiner that is the parent operation of the previous joiner.

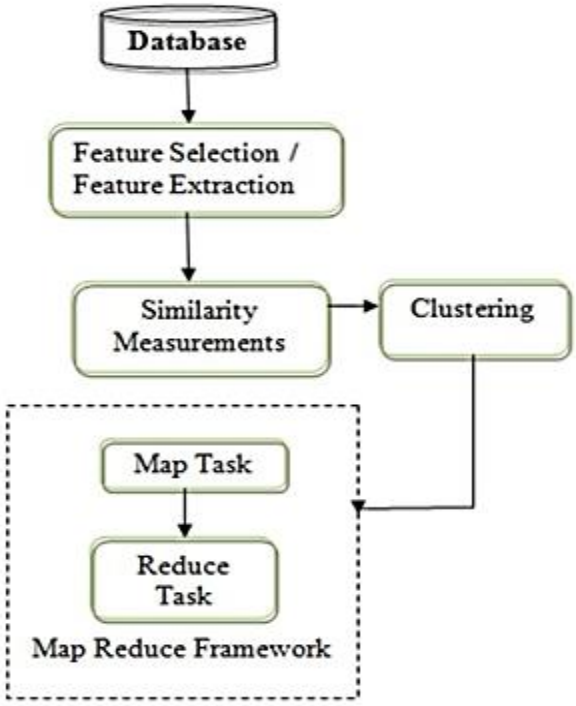

Fig.4. System framework

IV.CONCLUSION

We have demonstrated how to introduce various data clusters on map reduce. In this penetrate we also demonstrated how map reduce frameworks cooperate with various database management systems allowing for interesting possibilities. Efficiency is low with fault tolerance and scalability as its principle goals. Map reduce operations are not always allocated or computed for I/O efficient. For processing problems of large quantities of information map reduce answer is viable. Especially problems are partitioned into sub groups that can be worked out. Map reduce can become a most popular paradigm and popular solution. We have implemented about map reduce and it's specifications. Since, map reduce is so simple but it offers scalability its solves and manages massive information processing. Map reduce can be subsitute for DBMS and also for data warehousing.

V. REFERENCES

[1]. Indranil Palit and ChandranK.Reddy, Member, IEEE, scalable and parallel Boosting with MapReduce, IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24,NO.10, OCTOBER 2012.

[2]. Makho Ngazimbi, DATA CLUSTERING USING MAPREDUCE, March 2009.

[3]. K.E.Hemapriya,K.Gomathy,"A survey paper of cluster based key management techniques for secured data transmission in Manet". International Journal of Advanced Research in Computer and communication Engineering. (IJARCCE) vol 5, issue 10,October 2016.

[4]. TARUN DHAR DIWAN, PRADEEP CHOUKSEY, R. S. THAKUR & BHARAT LODHI, Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data, BIRT, Bhopal M. P. India.

[5]. Alina Ene, Sungjin Im, Benjamin Moseley, Fast Clustering using MapReduce.

[6]. Elena Tsiporkova, Veselka Boeva, Elena Kostadinova, MapReduce andfca Approach for Clustering of Multiple-Experiment Data.

Cite this article as : Kechika. S, Sapthika. B, Keerthana. B, Abinaya. S, Abdulfaiz. A , "A Systematic Algorithm for Data Cluster Using Map-Reduce Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 564-569, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195270
Journal URL : http://ijsrcseit.com/CSEIT195270