A Systematic Data Mining Method for Clustering of Data using Map-Reduce Model

J.Lekha, S.Kowsalya, S.Abinaya, B.Kiruthika, L.Nithya

Abstract: Data mining is an important research concept that has a vast scope in future. Data mining is used to find the unseen information from the data. In cluster, main half is feature choice. It involves recognition of a set of options of a set, because feature choice is taken into account as a necessary method. They additionally produce the approximate and according requests with the initial set of options employed in this kind of approach. The most construct on the far side this paper is to relinquish the end result of the bunch options. This paper conveys the cluster and the clustering process. The processing of large datasets the nature of clustering where some more concepts are more helpful and important in a clustering process. In clustering methodology many concepts are very useful. The feature selection algorithm which affects the entire process of clustering is the map-reduce concept. Here time needed to seek out the effective options, options of quality subsets is capable of providing effectiveness. The paper discussed map-reduce feature selection approach, its algorithm and framework of implementation.

Keywords : Data Mining, Clustering, Feature Selection, Map-Reduce, Map-Reduce Approach.

I. INTRODUCTION

Data mining could be a method of extraction of information and patterns from a collection of data. It is additionally known as data discovery method, data mining from knowledge, data extraction or knowledge /pattern analysis. Typically {data mining |data methoding} is that the process of partitioning a group of data (or objects) into a group of purposeful sub-classes [1,2,3]. Cluster is analogous to classification in that data are grouped. Cluster is not similar to classification in one way that in clusters, the groups are not pre-defined. The similarities of the data are found according to the characteristics found in the actual data. 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[2,3].

Whereas we introduce a map-reduce approach for clustering. In business, marketing purpose, cluster analysis is used to discover, implement and characterize customer segments.[3]

II. DATA PROCESSING TASKS

The data mining tasks may be classified typically into 2 varieties supported what a particular task tries to realize. Those two classes are square measure graphic tasks and extrapolative tasks. The expressive data processing tasks portray the final properties {of knowledge |of knowledge of information} whereas prognostic data processing tasks perform logical thinking on the offered knowledge set to expect however a brand new data set will behave. There square measure variety of knowledge mining tasks like classification, calculation, time-series analysis, suggestion, clustering, account etc. of these tasks square measure either prophetical data processing tasks or graphic data dealing out tasks. a knowledge mining system will complete one or supplementary of the on top of specific tasks as a part of data processing.[4,5,6].Systematic data processing tasks return up with a typical from the offered knowledge set that's useful in guessing unknown or future values of another knowledge set of interest.[2] A medical man attempting to diagnose a sickness supported the medical take a look at results of a persistent will bethought-about as a prophetical data processing task. A graphic data processing task typically find knowledge unfolding patterns and comes up with new, vital data from the offered knowledge set. A distributer attempting to spot product that square measure obtained along will be thought-about as a graphic data processing task[7].

a) Classification

A model is derived using classification to regulate the class of an object based on its attributes. A record has a set of characteristics and those collections of records will be available. Class characteristic is one of the attribute and the main task of classification is to assign a class attribute to a set of records. Classification will be utilized in marketing, that's to cut back selling prices by targeting a group of shoppers..Classification is useful in marketing tasks that is to minimize marketing costs by targeting a set of clients who are likely to buy a new produce. Using the presented data, it is possible to know which customers procured similar products and who did not purchase in the past[5,6].

Revised Manuscript Received on January 30, 2020.
* Correspondence Author

Dr.J.Lekha*, Assistant Professor, Department of Computer Science and Applications , Sri Krishna Arts and Science College, Coimbatore

Mrs.S.Kowsalya, Assistant Professor, Department of Computer Science and Applications ,Sri Krishna Arts and Science College, Coimbatore

S.Abinaya, Student, Department of Computer Science and Applications , Sri Krishna Arts and Science College, Coimbatore

B.Kiruthika, Student, Department of Computer Science and Applications , Sri Krishna Arts and Science College, Coimbatore

L.Nithyan, Student, Department of Computer Science and Applications , Sri Krishna Arts and Science, Coimbatore College

Hence, {purchase, don't purchase} conclusion forms the class attribute in this case. Once the class feature is assigned, demographic and lifestyle information of clients who bought similar products can be collected and upgrade mails can be sent to them straight.

b) Prediction

Prediction method calculates the attainable values of missing or future knowledge. Prediction involves evolving a model supported the offered knowledge and this model is employed in guessing future values of a replacement knowledge set of attention. for instance, a model will forecast the financial gain of Associate in Nursing workersupported education, expertise and differentde mographic issues like place of keep, gender etc. additionally prediction analysis employed in abundant areas as well as judgment, fraud detection etc.[3,4]

c) Time Series Analysis

Time sequence could be a arrangement of events wherever ensuing event is decided by one or additional of the previous events [7,8,9]. Statistic reflects method the method} being measured and there square measure bound parts that have an effect on the behaviors of a development. Statistic study includes strategies to research time-series knowledge so as to excerpt helpful outlines, trends, protocols and analytics. Stock exchange calculation is a vital claim of time- series investigation[5,7].

d) Association

Association identifies the association or affiliation amidst a group of things[6]. Association recognizes the relations between objects. Association investigation is employed for goods management, publicity, collection style, marketing etc. A distributer will decide the products that are frequently client's purchase together or analyze the clients who are attracted to the promotion of same kind of products. If a retailer identifies that beer and nappy are purchased together typically, he may put nappies on sale at a discount rate to encourage the promotion of beer[4,5,6].

e) Clustering

Clustering is used to recognize data objects that are like to one another. For sample, an insurance business can group its customers depending on age, address, salary etc.[5] This kind of data will be helpful to understand the customers better and hence offer better customized services.

f) Summarization

Summarization is the simplification of data. A group of data is made quick and that gives imperative material about the data. For example, the shopping done by a client can be concise into total products, total expenses, offers used, etc.[5,6,7] Such high level consolidated values can be useful for sales or customer connection team for completepurchaser and buying behavior analysis. Data can be grouped in a variety of thought levels and from changed angles.

III. CLUSTERING

Cluster is likely to classification in which data are combined together. Rather like classification, the groups are non-predefined. Similarly, grouping is accomplished by calculating similarities between data according to the characteristics of the actual data. We call these groups as 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[2,3,4]. Whereas we introduce a map-reduce approach for clustering. In business, marketing purpose, cluster analysis is used to discover, contrivance and illustrateclient segments. Clustering is used to predict data objects that are same like one another. For example, an insurance organization can cluster its customers based on age, address, salary etc[7]. This cluster data will be useful to recognize the clients better and so provides better modified services.

Limitations of clustering :

1. Programs occurring when clustering are tested with real world data store.
2. Outlier handling is difficult.
3. Dynamic data in the backend indicates the cluster association and may vary overtime.
4. Predicting the semantic meaning of each clusters may difficult. The exact meaning is of each cluster is may not clear. [7]
5. Another similar issue is which type of data must 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.[11,9]

Types of 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 demonstrate the parallel clustering technique and the groups of many types of clusters. The base in the histogram tree consists of a single cluster. The leaves in the tree consists of one element closer[4,5]. Base nodes in a tree represent new clusters formed by combining the clusters.

The levels in the tree is calculated with the relative distance measure that is applied to combine 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.[9,10]

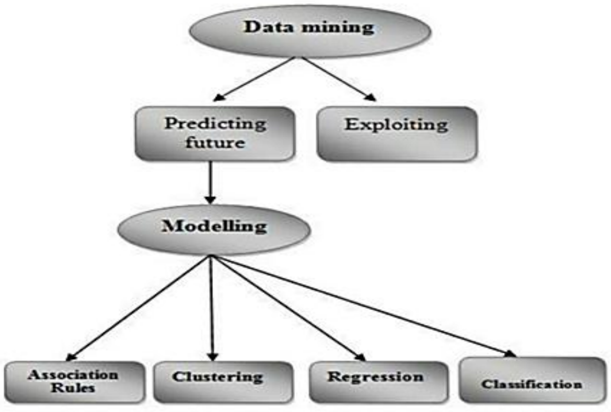

Fig.1.What is Data Mining& Classification

IV. 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.[9,12] 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 partitioned into four types: wrapping, hybrid, filter, and embedded method. Where wrapper method is used for optimization of the looping process.[11,12]

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.[1,3,4]

MAP Reduce Approach:

Map reduce step can solve a complex problem. Every steps takes an output from a previous step in map reduce.[1,2,3,4]

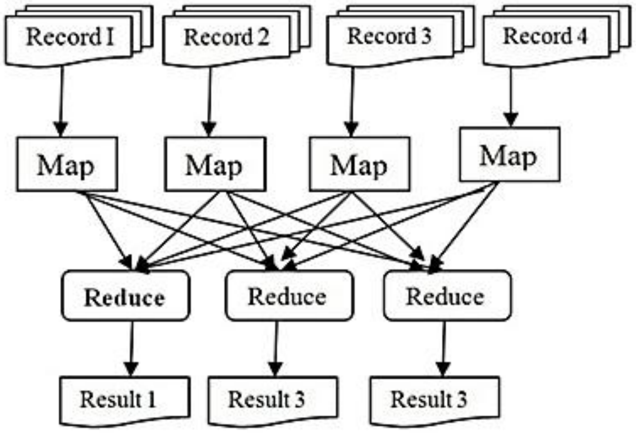

Fig.2.Map-Reduce Functionality

Data Preparation:

The collecting of huge amount of file each containing a data of one set is called data set. [14,15]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 can copy enter candidates 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. [1,5,6,7]

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 cennoter ID's. Its also removes duplication.

V. 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.[4,5,6] 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. [7,8,9]

Algorithm1:

```
MaRc algorithm with map reduce
Map function {(key value) /*(Record_id, Record_value),Tα*/}
Input: cluster data ( key, value)
Out put: 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: $P\alpha = f(key, T\alpha)$
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: $E\alpha = \{(key^\wedge, value)\ Pb\}$
4: for each (key^,value)€Eα}
5: $E\alpha = Tb\ n\ (key^\wedge, Pb)$ // 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) in between data to be reduce. [6,7,9]

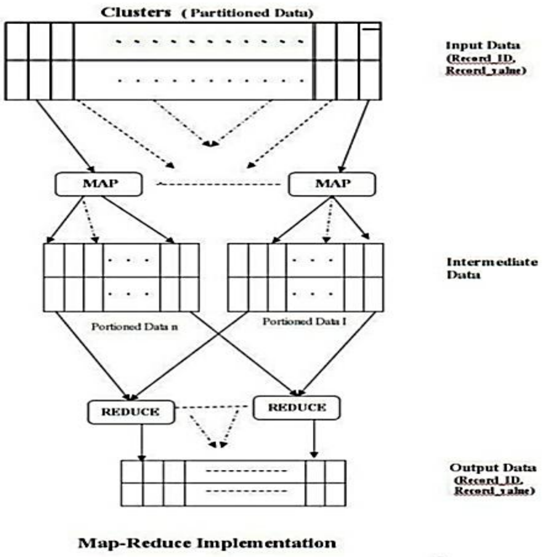
Clusters (Partitioned Data)
Input Data (Record_ID, Record_value)
MAP
MAP
Intermediate Data
Portioned Data n
Portioned Data I
REDUCE
REDUCE
Output Data (Record_ID, Record_value)
Map-Reduce Implementation

Fig.3. Map-Reduce Implementation

VI. MAP REDUCE FRAMEWORK

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. Joining and reducing actually which running inside the same reduce framework.[9,10,11] The map reduce join with two consecutive 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.[2,3,5]

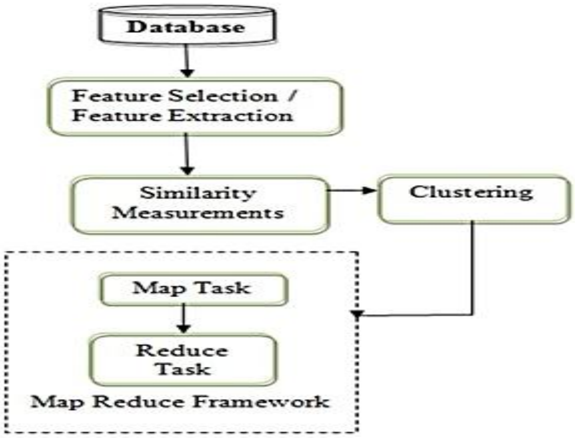

Fig.4. System framework

VII. 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[5,6,7]. 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. [5,6,7] 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. [4,5]

REFERENCES

1. IndranilPalit 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. MakhoNgazimbi, 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, SungjinIm, Benjamin Moseley, Fast Clustering using MapReduce.
6. 0Elena Tsiporkova, VeselkaBoeva, Elena Kostadinova, MapReduce and fca Approach for Clustering of Multiple-Experiment Data.
7. Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
8. Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005. International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010
9. Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.
10. Chapman, P., Clinton, J., Kerber, R., Khabaza, T.,Reinartz, T., Shearer, C. and Wirth, R.. "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".
11. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.
12. Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.
13. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process", Working Paper of the Center for Digital Economy Research, New York University and also presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.
14. Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," Proceedings of World Academy of Science, Engineering and Technology, April 2005.

AUTHOR'S PROFILE

Dr.J.Lekha, is currently working as an assistant professor in the department of Computer Science and Applications. She has 14 years of teaching experience. Her areas of specialization are network security, key management and data mining. She has guided 200 UG and PG projects and 5 M.Phil thesis. She has published more than 20 research articles in national, international journals and conferences indexed by Scopus, Web of Science and UGC databases.

Mrs.S.Kowsalya, is working as Assistant Professor in Department of Computer Science and Applications at Sri Krishna Arts and Science College. She Received M.Phil from Bharathiyar University in 2015 and her research interest span in Data Mining.

S.Abinaya, is pursuing her final year B.Sc CSA at Sri Krishna Arts and Science College. Her areas of interest include database languages and data mining.

B. Kiruthika, is pursuing her final year B.Sc CSA at Sri Krishna Arts and Science College. Her areas of interest include programming languages and data mining.

L.Nithyan, is pursuing his final year B.Sc CSA at Sri Krishna Arts and Science College. Her areas of interest include database languages and data mining.