Enhancement of Clustering Mechanism in Grid Based Data Mining

Ritu Devi
M.Tech student, Department of CSE,
Jind Institute of Engineering and Technology, Jind (Haryana)

Gurdev Singh
Assistant Professor, Department of CSE,
Jind Institute of Engineering and Technology, Jind (Haryana)

Abstract: Distributed data mining (DDM) techniques have become necessary for large and multi-scenario datasets requiring resources, which are heterogeneous and distributed in nature. We focus our attention on distributed data mining approach via grid. We have discussed and analyzed a new framework based on grid environments to execute new distributed data mining approaches that best suits a distributed and heterogeneous datasets that are commercially available. The architecture and motivation for the design have also been presented in this paper. A detailed survey on distributed data mining technology was also carried out which could offer a better solution since they are designed to work in a grid environment by paying careful attention to the computing and communication resources.

I. INTRODUCTION

Distributed data mining is originated from need of mining over decentralized data sources. Field of Distributed Data Mining (DDM) deals with these challenges in analyzing distributed data & offers many algorithmic solutions to perform different data analysis & mining operations in a fundamentally distributed manner that pays careful attention to resource constraints. It surveyed data mining literature on distributed & privacy-preserving clustering algorithms. It discussed sensor networks with peer-to-peer architectures as an interesting application domain & illustrated some of existing challenges & weaknesses of DDM algorithms. It noted that while these algorithms usually perform better than their centralized counter-parts on grounds of communication efficiency & power consumption, there exist several open issues. Grid computing, simple stated, is taking distributed computing to the next level. So firstly a short definition of distributed computing followed by the definition of grid computing. Distributed computing means dividing tasks among multiple computer systems instead of doing the tasks on one centralized computer system. Distributed computing is a subset of grid computing, grid computing encompasses much more.

II. DISTRIBUTED APPROACH IN A DATABASE

Data mining had attracted a great deal of attention in information industry & in society as a whole in recent years, due to wide availability of huge amounts of data & imminent need for turning such data into useful information & knowledge. Data mining refers to extracting or "mining" knowledge from large amounts of data. Distributed data mining (DDM) is data mining where data & computation are spread over many independent sites. Each site had its own data source & data mining algorithms producing local models. From them global meaningful knowledge had to be derived.

III. DESIGN METHODOLOGY

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful & understandable patterns in large databases. Patterns must be actionable so that they may be used in an enterprise's decision making process. It is usually used by business intelligence organizations, & financial analysts, but this is increasingly used in sciences to extract information from enormous data sets generated by modern experimental & observational methods.

A typical example for a data mining scenario may be "In context of a super market, if a mining analysis observes that people who buy pen tend to buy pencil too, then for better business results seller could place pens & pencils together."

Data mining strategies could be grouped as follows:

3.1 Classification

Here given data instance has to be classified into one of target classes which are already known or defined [19, 20]. One of examples could be whether a customer has to be classified as a trustworthy customer or a defaulter with in a credit card transaction data base, given his various demographic & previous purchase Characteristics.

3.2 Estimation

Like classification, purpose of an estimation model is to determine a value for an unknown output attribute. However, unlike classification, output attribute for an estimation problem are numeric rather than categorical. An example could be "Estimate salary of an individual who owns a sports car?"

3.3 Prediction

It is not easy to differentiate prediction from classification or estimation. Only difference is that rather than determining current behavior, predictive model predicts a future outcome. Output attribute could be categorical or numeric.

3.4 Association rule mining

Here interesting hidden rules called association rules with in a large transactional data base is mined out. For e.g. rule {milk, butter → biscuit} provides information that whenever milk & butter are purchased together biscuit is also purchased, such that these items could be placed together for sales to increase overall sales of each of items.

3.5 Clustering

Clustering is a special type of classification with in which target classes are unknown. For e.g. given 100 customers they have to be classified based on certain similarity criteria & this is not preconceived which are those classes to which customers should finally be grouped into.

Here we are using clustering strategy.

Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Help users understand natural grouping or structure with in a data set. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

Cluster analysis or clustering is task of grouping a set of objects with in such a way that objects within same group (called a cluster) are more similar (in some sense or another) to each other than to those within other groups (clusters). It is a main task of exploratory data mining, & a common technique for statistical data analysis, used within many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, & computer graphics. Cluster analysis itself is not one specific algorithm, but general task to be solved. It could be achieved by various algorithms that differ significantly within their notion of what constitutes a cluster & how to efficiently find them. Popular notions of clusters include groups with small distances among cluster members, dense areas of data space, intervals or particular statistical distributions. Clustering could therefore be formulated as a multi-objective optimization problem. Appropriate clustering algorithm & parameter settings (including values such as distance function to use, a density threshold or number of expected clusters) depend on individual data set & intended use of results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial & failure. It is often necessary to modify data preprocessing & model parameters until result achieves desired properties.

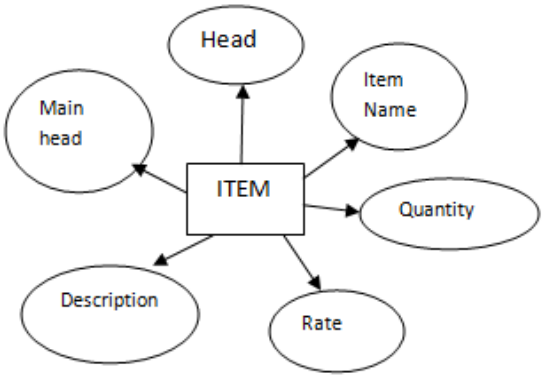
IV. GRID COMPUTING

The idea behind grid computing is that one can plug one's computer into the wall and have access to computational and data resources without knowing where they are or who owns these resources. Distributed computing is a subset of grid computing, grid computing encompasses much more. Grid computing provides coordinated sharing of geographically distributed hardware, software and information resources, this sharing is highly controlled defining clearly what is shared, who is sharing and the conditions of the sharing, it provides a service oriented infrastructure and uses standardized protocols to accomplish this sharing.

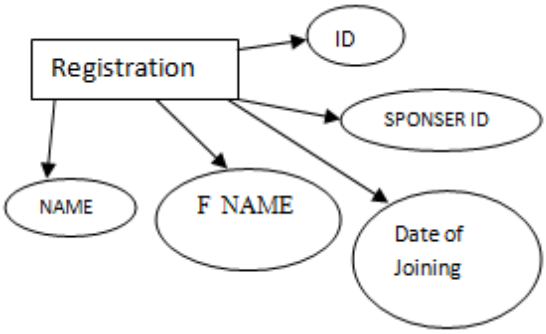
V. IMPLEMENTATION

Dataset description this research we deal with dataset of Sugar mill and MLM company database. Here we have taken average data of two databases, one is opening quantity of items used in Sugar mill. Second, we have taken opening quantity of id and sponser_id in MLM Company.

Entity relationship model of Sugar mill



Entity relationship model of MLM



Comparative analysis between grid & non grid based platform

Table1. Grid & non grid based platform

Number of record	Time (Non- Grid Based)	Time (Grid Based)
1000	20	5
2000	36	9
3000	57	15
4000	78	20
5000	90	23
6000	109	27
7000	131	33
8000	155	40
9000	170	43
10000	187	47

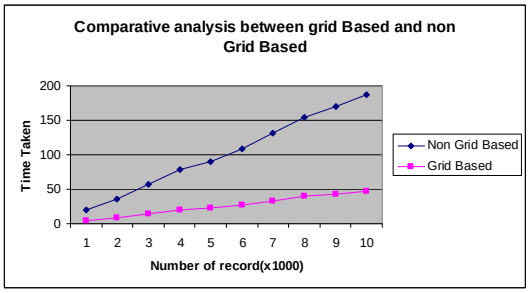


Fig1. Comparison between Grid & non- grid based

Table2. Different cases with different number of item in every data set the cluster were generated according to number of records in file

Test	Records	Number of Cluster Generated
1	5000	5
2	3200	4
3	2034	3
4	1500	2
5	500	1
6	400	1
7	200	1
8	100	1
9	50	1

Table3. Number of cluster

Test	Records	Number of clusters generated	Non-Empty Clusters	Empty Clusters
1	5000	5	4	1
2	3200	4	3	1
3	2034	3	2	1
4	1500	2	2	0
5	500	1	1	0
6	400	1	1	0
7	200	1	1	0
8	100	1	1	0
9	50	1	1	0

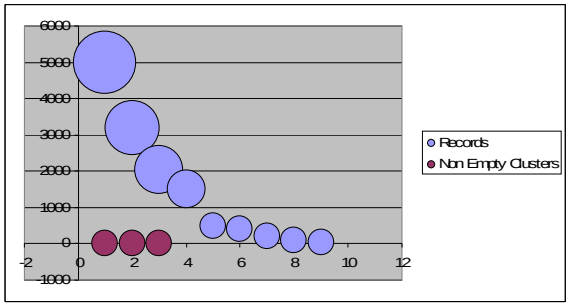


Fig2. Empty clusters were generated according to the distance among data items the creation of empty cluster depends on data items.

Output in case of old K-Means:

Table4. Number of record in cluster and total size of cluster in both cases

Number of record	No. of Cluster in case of old k-means	Total Size in old k-means	No. of clusters in case of new k-means	Total size in new k-means
1000	2	1220	1	1123
2000	3	1843	2	1750
3000	4	2490	3	2276
4000	6	4945	4	4760
5000	8	6734	5	6593
6000	9	7554	6	7345
7000	9	8454	7	8322
8000	12	12344	8	12222
9000	13	13454	9	12954
10000	14	15667	10	14322

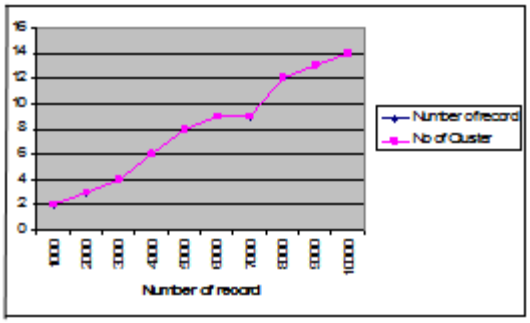


Fig3. Number of record & cluster in old k-means

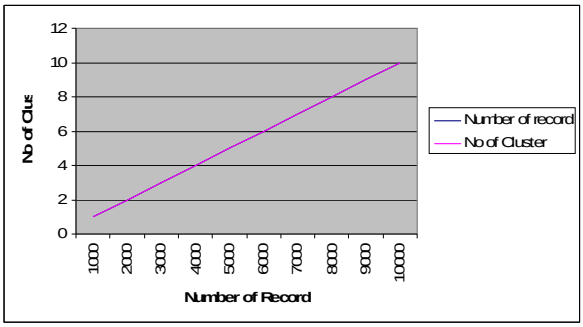


Fig4. Number of record & cluster in new k-means

Comparative analysis of result between old & enhanced K-MEAN

Table5. Comparative analysis of result between old & enhanced K-MEAN

Number of record	Old K-Mean Algorithm	Enhanced Algorithm
1000	2	1
1500	2	2
2000	3	2
2500	3	3
3000	4	3
4000	6	4
5000	8	5
6000	9	6
7000	9	7
8000	12	8
9000	13	9
10000	14	10

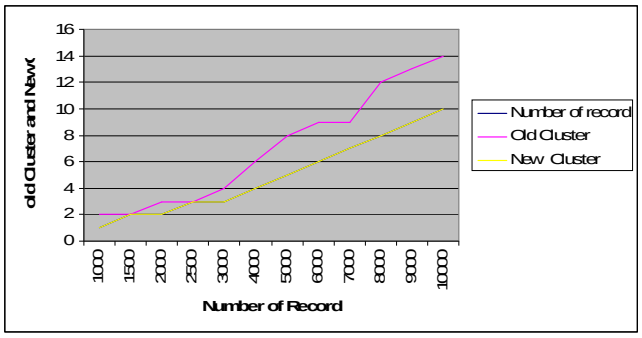


Fig5. Analysis of old & new cluster

Above figure represent comparative analysis of number of clusters formed in case of old K mean clustering & enhanced K mean clustering. Number of vacant clusters has been removed in case of enhanced clustering algorithm so number of clusters got reduced in case of enhanced algorithm.

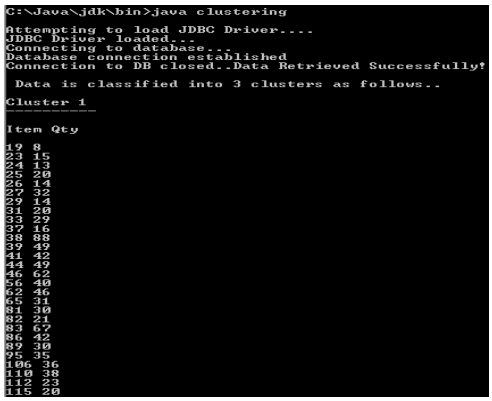


Fig6. Item in cluster 1

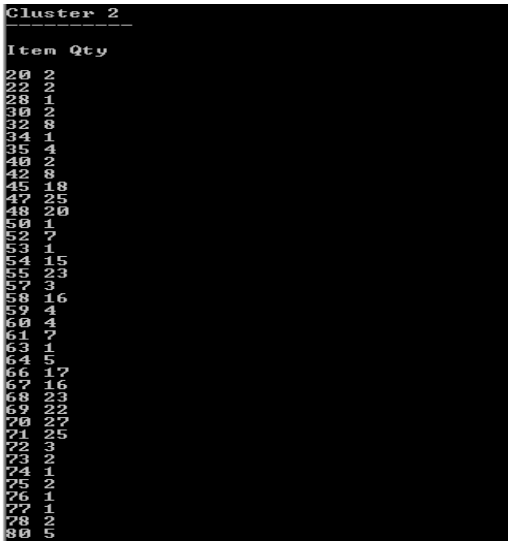


Fig7. Item in cluster 2

VI. CONCLUSION AND FUTURE SCOPE

Distributed data mining has been influenced from decentralized data sources. In this research we have reduced Number of vacant clusters so number of clusters gets reduced in case of enhanced algorithm.

Distributed Data Mining usually deals with challenges in analyzing distributed data. It also offers several algorithmic solutions in order to perform different data analysis. Here we have also solved clustering issues in case of traditional K-Mean Algorithm. In grid computing one can plug one's computer into the wall and have access to computational. In this research Comparative analysis between grid & non grid based has been done. The performance gets better in grid based computing as compared to non grid based computing. The advent of laptops, palmtops, cell phones, and wearable computers is making ubiquitous access to large quantity of data possible. Advanced analysis of data for extracting useful knowledge is the next natural step in the world of ubiquitous computing. Accessing and analyzing data from a ubiquitous computing device offer many challenges.

REFERENCES

[1]. J. Liu, S. Zhang, Y. Ye, Agent-based characterization of web regularities, in N. Zhong, et al. (eds.), Web Intelligence, NewYork: Springer, 2003, pp. 19–36.

[2]. J. Liu, N. Zhong, Y. Y. Yao, Z. W. Ras, wisdom web: new challenges for web intelligence (WI), J. Intell. Inform. Sys.,20(1): 5–9, 2003.

[3]. Congiusta, A. Pugliese, D. Talia, & P. Trunfio, Designing GridServices for distributed knowledge discovery, Web Intell. Agent Sys, 1(2): 91–104, 2003.

[4]. J. A. Hendler & E. A. Feigenbaum, Knowledge is power: semantic web vision, in N. Zhong, et al. (eds.), Web Intelligence: Research & Development, LNAI 2198, Springer, 2001, 18–29.

[5]. N. Zhong & J. Liu (eds.), Intelligent Technologies for Information Analysis, New York: Springer, 2004.
[6]. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project".
[7]. Journal of Machine Learning Research **11**: 2533–2541. Original title, "Practical machine learning", was changed ... term "data mining" was [added] primarily for marketing reasons.
[8]. Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, & Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.
[9]. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, & Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012.
[10]. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, & Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
[11]. "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.
[12]. "Google Scholar: Top publications - Data Mining & Analysis". Google Scholar.
[13]. Proceedings, International Conferences on Knowledge Discovery & Data Mining, ACM, New York.