IJCSC 0973-7391 Volume 8 • Issue 1 Sept 2016 - March 2017 pp. 1-3 available online at www.csjournalss.com

Review on Enhancement of Clustering Mechanism in Grid Based Data Mining

Ms Ritu Devi[1], Gurdev Singh[2]
[1]M.Tech student, Department of CSE, Jind Institute of Engineering and Technology, Jind (Haryana)
[2]Assistant Professor, Department of CSE, Jind Institute of Engineering and Technology, Jind (Haryana)

Abstract: Today organizations often use data from several resources. Data is characterized to be heterogeneous, unstructured & usually involves a huge amount of records. This implies that data must be transformed in a set of clusters, parts, rules or different kind of formulae, which helps to understand exact information. Participation of several organizations in this process makes assimilation of data more difficult. Data mining is a widely used approach for transformation of data to useful patterns, aiding comprehensive knowledge of concrete domain information. Nevertheless, traditional data mining techniques find difficulties in their application on current scenarios, due to complexity previously mentioned. Data Mining Grid tries to fix these problems, allowing data mining process to be deployed in a grid environment, in which data & services, resources are geographically distributed belong to several virtual organizations & security can be flexibly solved. We propose both a novel architecture for Data Mining Grid, named DMG.
Keywords: DMG, Clustering, K-Mean, SOM, EM clustering, WAN.

1. INTRODUCTION

Data Mining refers to process of extracting useful, handy & survivable knowledge from data. Extracted knowledge is useful in many areas such as business applications like financial business analysis, purchasing behavior scenarios & also in biology, molecular design, weather forecast, climate prediction, physics, fluid dynamics & so on. Now challenge in these applications is to mine data located in distributed, heterogeneous databases while adhering to varying security & privacy constraints imposed on local data sources. Term grid can be defined as a set of computational resources interconnected through a WAN, aimed at performing highly demanding computational tasks such as in internet applications. A grid makes it possible to securely & reliably take advantage of widely dispersed computational resources across several organizations & administrative domains. Aim of grid computing is to provide an affordable approach to large-scale computing problems. Grid technology provides high availability of resources & services, making it possible to deal with new & more complex problems. But it is also known that a grid is a very heterogeneous & decentralized environment [8]. It presents different kinds of security policy, system administration procedure, data & computing characteristic & so on. In this juncture we can't say that any grid is not just a data mining grid. It is a very important aspect in maintaining grid systems. Grid management is the key to providing high reliability & quality of service [9]. Complexities of grid computing environments make impossible to have a complete understanding of entire grid. Therefore, a new approach is needed. Such an approach should pool, analyze & involve all relevant information that could be obtained from a grid. Insights provided should then be used to support resource management & system involvement.

2. DATAMINING WITH GRID

Data mining is a widely used approach for transformation of data to useful patterns, aiding comprehensive knowledge of concrete domain information. Nevertheless, traditional data mining techniques find difficulties in their application on current scenarios, due to complexity previously mentioned. Data Mining Grid tries to fix these problems, allowing data mining process to be deployed in a grid environment, in which data & services resources are geographically distributed belong to several virtual organizations & security could be flexibly solved. We propose both a novel architecture for Data Mining Grid, named DMG. Data Mining Grid needs frequently exchange of data mining models among participating sites. Therefore, seamless & transparent realizations of DMG technology would require standardize schemes to represent & exchange models.

3. K-MEANS CLUSTERING ALGORITHM

Clustering

Clustering analysis [1] is broadly used in many applications such as market research, style recognition, data analysis, & image processing. Clustering could also help marketers discover [1] distinct groups in their customer base. & they could characterize their customer groups based on purchasing patterns. A cluster of data objects could be treated as one group.

While doing cluster analysis [1], we first partition set of data into groups based on data similarity & then assign labels to groups. Main advantage of clustering over classification [7]

DOI: 10.031206/IJCSC.2016.001

is that, it is adaptable to changes & helps single out useful features that distinguish different groups.

4. APPLICATION OF CLUSTER ANALYSIS

The cluster analysis [1] has been applied to many occasions. For example, in commercial, cluster analysis was used to find different customer groups, & summarize different customer group characteristics through buying habits; cluster analysis was used to categorized animal & plant populations according to population & to obtain latent structure of knowledge; in geography, clustering could help biologists to determinate relationship of different species & different geographical[6] climate; in banking sector, by using cluster analysis to bank customers to refine a user group; in insurance industry, according to type of residence, around business district, geographical[14] location, cluster analysis could be used to complete an automatic grouping of regional real estate, to reduce manpower cost & insurance company industry risk; in Internet, cluster analysis was used for document classification[8] & information retrieval etc.

Types of Clustering Algorithms are:

1. K-means Clustering Algorithm
2. Hierarchical Clustering Algorithm
3. Density Based Clustering Algorithm
4. Self-organization maps (SOM)
5. EM clustering Algorithm

5. CLUSTERING PROCESS

The clustering process consists of following steps

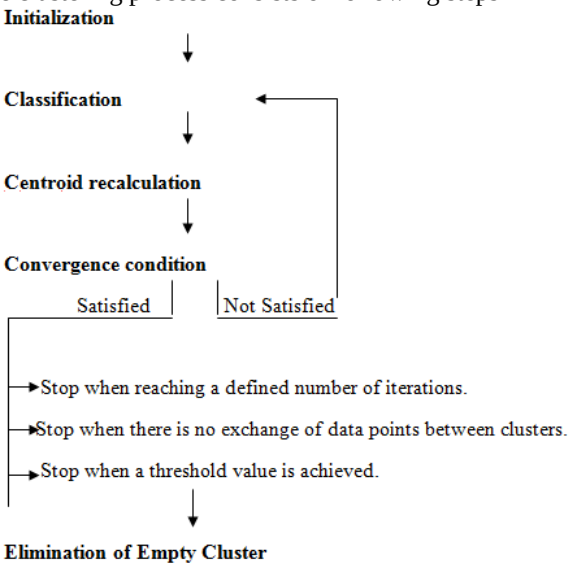

Fig 1 New K-Mean based algorithm

1. **Initialization:** In this first step data set, number of clusters & centroid should be calculated automatically according to size of data.

2. **Classification:** distance is calculated for each data point from centroid & data point having minimum distance from centriod of a cluster is assigned to that particular cluster.

3. **Centroid Recalculation:** Clusters generated previously, centriod is again repeatedly calculated means recalculation of centriod.

4. **Convergence Condition:** Some convergence conditions are given as below:

   I. Stopping when reaching a given or defined number of iterations.
   II. Stopping when there is no exchange of data points between clusters.
   III. Stopping when a threshold value is achieved.

5. **Take steps:** If all of above conditions are not satisfied, then go to step 2 & whole process repeated again, until given conditions are not satisfied.

6. **Elimination of Empty Clusters:** *Clusters generated previously are rechecked*

Clusters where no data points are allocated to a cluster under consideration during assignment phase are eliminated.

Benefits of new algorithm over traditional

i. No need of predefined cluster center

ii. There would be no Empty clusters at end

6. CONCLUSION AND FUTURE SCOPE

Clustering is process [14] of grouping objects that belongs to same class. Similar objects are grouped in one cluster & dissimilar objects are grouped in another cluster. Clustering analysis is used in several applications like market research, pattern recognition, data analysis [1]. K-means clustering is very fast, robust & easily understandable. If data set is separated from one other data set, then it gives best results. Clusters do not having overlapping character & are also non-hierarchical within nature. One more problems[6] with K-means clustering is that empty clusters are generated during execution[14], if within case no data points are allocated to a cluster under consideration during assignment phase. Our work is to remove empty cluster and do automatic clustering.

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. Since multi-agent systems are often distributed & agents have proactive & reactive features which are very useful for Knowledge Management Systems, combining DDM with MAS for data intensive applications is appealing. This paper integration of grid system & distributed data mining, also known as grid based

DOI: 10.031206/IJCSC.2016.001

distributed data mining, in terms of significance, system overview, existing systems, & research trends.

Due to increasing amount of data available online, World Wide Web has becoming one of the most valuable resources for information retrievals & knowledge discoveries.

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.

DOI: 10.031206/IJCSC.2016.001