IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 6, Ver. II (Nov – Dec. 2015), PP 53-54
www.iosrjournals.org

A Hierarchical and Grid Based Clustering Method for Distributed Systems (Hgd Cluster)

Mr. A.Venugopal M.Sc., M.Phil, V.Sujatha Veni.
[1]*Assistant Professor Sree Narayana Guru College Coimbatore*
[2]*Research Scholar Sree Narayana Guru College Coimbatore.*

Abstract: *In distributed peer-to-peer systems, huge amount of data are dispersed. Grouping of those data from multiple sources is a tedious task. By applying effective data mining techniques the clustering of distributed data is become ease and this decreases the hurdles of clustering due to processing, storage, and transmission costs. To perform a dynamic distributed clustering, a fully decentralized clustering method has been proposed. HGD Cluster can cluster a data set which is dispersed among a large number of nodes in a distributed environment using hierarchal and grid based clustering techniques. When nodes are fully asynchronous and decentralized and also adaptable to stir, then HGD cluster will apply. The general design principles employed in the proposed algorithm also allow customization for other classes of clustering. It is fully capable of clustering dynamic and distributed data sets. Using the algorithm, every node can maintain summarized views of the dataset. Customizing HGD Cluster for execution of the hierarchal-based and grid-based clustering methods on the summarized views is the main aim of the proposed system. Coping with dynamic data is made possible by gradually adapting the clustering model.*

I. Introduction

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing strategies for their product. They can use data to compare and contrast among competitors. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company.

II. Heading

1. Dataset collection
2. Group pattern analysis
3. Group pattern analysis
4. Density calculation
5. Hash index maintenance
6. Bloom search option.

III. Indentation and Equation

1. Dataset collection

The system utilizes dynamic datasets, where P2P groups are dynamic in nature and distributed systems change continuously, because of nodes joining and leaving the system, or because their set of internal data is modified. So the proposed system has the ability to handle dynamic data environment. So the initial process is to initiate the process with dataset collection.

2. Group pattern analysis

The goal of this phase is to propose an efficient data mining mechanism for deriving object group prototypes and utilize the object tracking patterns for fast prediction. To facilitate collaborative data collection

processing in object tracking P2P networks, cluster architectures are usually used to organize P2P nodes into clusters.

3. Implementation of HGD

The third phase implements the hierarchical grid network for fast data summarization and search. This has the following process

- Creation of the Grid Structure
- Calculation of the block density
- Sorting of the blocks
- Identifying cluster centers

4. Density calculation

5. Hash index maintenance

Traversal of neighbour blocks and stores the results in hash index using bloom search.

6. Bloom search option.

IV. Conclusion

A summary and data search over distributed data is more tedious, to simply the task of data management over P2P, this introduced a new HGD cluster technique, which aims to identify and track the movement of objects. It provided the effective tracking mechanism in a distributed clustering environment. To address the group data management and summary problems in the distributed environment, the proposed system has discovered Grid and hierarchical clustering with ensemble algorithm. HGD first identified the necessity of an effective and efficient distributed clustering algorithm. Dynamic nature of data demands a continuously running algorithm which can update the clustering model efficiently, and at a reasonable pace. This improved existing GDCluster with different clustering mechanism named as HGDCluster a general fully decentralized clustering algorithm, and instantiated it for Grid-based and Hierarchical-based clustering methods. The proposed algorithm enabled nodes to gradually build a summarized view on the global data set hierarchical, and execute weighted clustering algorithms to build the clustering models. It reduces the cost of data management and summarization process over distributed environment. This utilizes the previous search results for fast data summarization, so this used bloom search mechanism. Finally the system shows the comparative results and efficiency of the proposed system.

Reference

[1] Xiong, Sicheng, Javad Azimi, and Xiaoli Z. Fern. "Active learning of constraints for semi-supervised clustering." Knowledge and Data Engineering, IEEE Transactions on 26.1 (2014): 43-54.

[2] Basu, Sugato, Arindam Banerjee, and Raymond J. Mooney. "Active Semi-Supervision for Pairwise Constrained Clustering." SDM. Vol. 4. 2004.

[3] Bilenko, Mikhail, Sugato Basu, and Raymond J. Mooney. "Integrating constraints and metric learning in semi-supervised clustering." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[4] Davidson, Ian, Kiri L. Wagstaff, and Sugato Basu. Measuring constraint-set utility for partitional clustering algorithms. Springer Berlin Heidelberg, 2006.

[5] Greene, Derek, and Pádraig Cunningham. "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering." Machine Learning: ECML 2007. Springer Berlin Heidelberg, 2007. 140-151.

[6] Guo, Yuhong, and Dale Schuurmans. "Discriminative batch mode active learning." Advances in neural information processing systems. 2008.

[7] S. Basu, A. Banerjee, and R. Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering," Proc. SIAM Int'l Conf. Data Mining, pp. 333-344, 2004.

[8] P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering," Proc. Int'l Conf. Pattern Recognition, pp. 1-4, 2008.

[9] D. Greene and P. Cunningham, "Constraint Selection by Committee: An Ensemble Approach to Identifying Informative Constraints for Semi-Supervised Clustering," Proc. 18th European Conf. Machine Learning, pp. 140-151, 2007.

[10] M. Al-Razgan and C. Domeniconi, "Clustering Ensembles with Active Constraints," Applications of Supervised and Unsupervised Ensemble Methods, pp. 175-189, Springer, 2009.