Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

A Review on Association rule mining in distributed environment using DIC algorithm

Patel Kartik K[1], Tanvi Varma[2]
[1,2]*Department of Computer Science and Engineering,*
Parul Institute of technology

Abstract— Today associations are artificially distributed. Typically, every site locally stores its daily Updated data. Using centralized data mining to invention of efficient patterns in such organizations data does not every time potential because combining datasets from different sites into a centralized location because of broad network communication costs. If it is impossible to combine them in a central location. Distributed data mining has an active subarea of data mining. In distributed association rule mining algorithm, one of the major challenges is to commute the communication overhead. Data sites are required to exchange of information in the data mining process which may generates communication overhead. A challenge is to commute number of database scan and generate the frequent itemsets from the database.

Keywords- Data Mining, Association Rule Mining, Distributed Data Mining, Centralized Data Mining, Distributed Association Rule.

I. INTRODUCTION

The explosive growth of the amount of data gathered by transactional systems, a challenge for finding new techniques to extract useful patterns from such a huge amount of data arose. Data mining emerged as the new research area to meet this challenge. Data mining is one of the means to utilize information by discovering underlying hidden useful knowledge from information [2]. Technology advances have made data collection easier and faster, resulting in large, more complex, datasets with many objects and dimensions. Important information is hidden in this data. Data Mining has become an intriguing and interesting topic for the information extraction from such data collection since the past decade. Furthermore there are so many subtopics related to it that research has become a fascination for data miners.

Association rule mining is one of the popular techniques for mining data [4]. In this technique, an interrelation among different items in data is discovered by determining frequent large item-sets which are repeated more than a threshold number of times in the database [2].

Data mining process can be characterized as centralized and distributed based on the location of data [2].In centralized data mining data is stored on single site and the main purpose of the efficiency of a data mining algorithm and its I/O and CPU time. The I/O time is the number of disk reads or the number of passes of the database made by the algorithm [1].And in Distributed data mining the data is resided into multiple sites. The data may be owned by each site separately or an enormous amount of data may be distributed into multiple data sites [2].To minimizes communication overhead and reduce number of database scan the algorithms are requires in distributed environment.

II. LITERATURE SURVEY

Preeti Paranjape, Umesh Deshpande[1], "An Optimistic Messaging Distributed Algorithm for Association Rule Mining" in this title distributed algorithm based on Dynamic Itemset Counting (DIC) for generation of frequent itemsets [1]. The nature of DIC represents a higher number of passes of the database and the total amount of time taken to obtain the frequent itemsets is reduced as compared to Apriori-based algorithms. In the proposed Optimistic messaging DIC focuses on disk I/O minimization by reducing the number of database passes and has almost no issue of synchronization between the nodes [1].

Md. Golam Kaosar, Zhuojia Xu and Xun Yi, "Distributed Association Rule Mining with Minimum Communication Overhead" Today in distributed environment one of the major and challenging task is to reduce the communication overhead [2]. In this paper propose an association rule mining algorithm which minimizes the communication overhead among the participating data sites [2]. Instead of transmitting all itemsets and their counts, they propose to transmit a binary vector and count of only frequently large itemsets [2]. Message Passing Interface (MPI) technique is exploited to avoid broadcasting among data sites [2].

Hadj-Tayeb karima, Hadj-Tayeb Lilia, "Distributed Data Mining by associated rules: Improvement of the Count Distribution algorithm" Today large system has overwhelmed by inundation of data that are store daily in distributed system. This paper present contributions to improve the algorithm by reducing the number of exchanged messages, and the number of generated candidates. The results showed that the proposed algorithm meets the expected objectives by presenting a performance gain greater than the CD algorithm in which the last points are important performance factors in determining the quality of an algorithm for extraction rules [3].

III. VARIOUS METHODS IN DISRTIBUTED ASSOCIATION RULE MINING

Various algorithms for generation of large frequent itemsets are presented as follows:

3.1 Count distribution algorithm

Count distribution is well known parallel interpretation of the sequential apriori algorithm. This algorithm partitions and distributed horizontally and equitably the database in all processors [3]. This algorithm is suitable for a model in which the computational capacities distinguish the communicational capabilities but it suffers from heavy communication cost at the broadcast of generated candidates. Count distribution transmits the count of itemsets since all sites have identical set of itemsets. It reduced the amount of overhead to be transmitted in network [2].This algorithm suffer the raise number of nodes and the size of the databases.

Advantages:

Improve the reducing the number of exchanged messages and performance gain and extracting frequent related itemsets based on the parallelism of the task.

3.2 Fast Distribution Mining Algorithm

The researchers mine the rule from distributed data from different sites. Fast distribution algorithm minimizes the number of candidates that are created using two pruning techniques: global and local pruning [3].In FDM local pruning technique use to reduces the number of items. The local pruning removed the element Y of all candidates if Y is not locally frequent. Then after this done each site add received local support to generate the globally frequent itemsets, this techniques is called global pruning. FDM reduce the number of messages sent between sites [3].

Advantages:

This algorithm tacit to reduce the number of candidates and reduced the communication overhead.

3.3 Dynamic itemset counting

Dynamic itemsetcounting is based on distributed algorithm for generation of frequent itemsets. DIC Algorithm reduced the number of passes made over the data while keeping the number of itemsets which are counted in any pass relatively low [1].DIC is does not wait for total database pass to start counting the candidates itemsets.It reduced the number of passes of database and generate less number of candiadate itemsets.

Advantages:

It gives a relationship between focus and non-focused motion of the objects.

3.4 Optimistic dynamic itemset counting (OPT-DIC)

OPT-DIC is a distributed version of dynamic itemset counting .In DIC performance affect the distributed environment are disk I/O minimization. This algorithm overcomes the synchronization time between nodes and transmission of messages over the network. OPT-DIC runs DIC at each node. DIC reads transactions and perform all operations of incrementation of the counters and adding supersets of items which become frequent. In OPT-DIC algorithm we also send and receive messages at the checkpoint [1]. At each checkpoint, every incoming queue is checked for counts of items which become potentially frequent from other sites [1].

Advantages:

Amount of time taken improve and reduction in the number of passes of the database and comparatively less number of candidate generated.

IV. CONCLUSION

In this paper, various techniques for frequent pattern itemset generation are explained. Using count distribution algorithm endures the increased number of nodes and the size of the database. Fast Distribution algorithm introduces some techniques to minimize candidate itemsets, it overloads the network by broadcasting too much data. Using OPT-DIC we can derive better result than count distribution in higher performance gain and dense dataset.

ACKNOWLEDGMENT

With the cooperation of my guide, I am highly indebted to Asst. Prof. TANVI VARMA, for her valuable guidance and supervision regarding my topic as well as for providing necessary information regarding review paper.

REFERENCES

[1] Preeti Paranjape, Umesh Deshpande, “An Optimistic Messaging Distributed Algorithm for Association Rule Mining”, IEEE-Dec 2009.
[2] Md. Golam Kaosar, Zhuojia Xu and Xun Yi, “Distributed Association Rule Mining with Minimum Communication Overhead”, IEEE-Jan 2009.
[3] Preeti Paranjape , Umesh Deshpande, “Research On Distributed Mining Algorithm For Association Rules Oriented Mass Data”, Proceedings of the 33rd Chinese Control Conference,July 28-30, IEEE-2014.
[4] Nidhi Sethi1 and Pradeep Sharma, “Mining Frequent Pattern from Large Dynamic Database Using Compacting Data Sets”, IJSRCSE-2013.
[5] Ms. Vinaya sawant,Dr. Ketan shah, “a review of distributed data mining using agents”, IJATER, 2013.
[6] Hadj-Tayeb karima, Hadj-Tayeb Lilia, “Distributed Data Mining by associated rules: Improvement of the Count Distribution algorithm”, IJCSI-May 2012.
[7] Sergey Brin,Rajeev motwani,jeffrey d ullman,shalom tsur, “Dynamic itemset counting and Implication rules for market basket analysis”, SIGMOD Record volume 6,1997.
[8] R.Agrawal and J.Schafer, “Parallel Mining of Association Rules”, IEEE Transactions on Knowledge and Data Engineering, 1996.

