Review Article | Impact Factor: 0.621 | ISSN: 2319-507X
Akhilesh Bhuyar, IJPRET, 2014; Volume 2 (9): 1025-1030 | IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

"PERFORMANCE IMPROVEMENT USING CLUSTERING IN DISTRIBUTED DATABASE"

MR. AKHILESH B. BHUYAR[1], DR. H. R. DESHMUKH[2], MS. R. N. GADBAIL[3], MR. R. G. ANANTWAR[3]

1. M. E. First Year, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.
2. Prof and HOD, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.
3. Asst. Prof, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: Association rules mining (ARM) algorithms have been extensively researched in the last decade. Therefore, numerous algorithms were proposed to discover frequent item sets and then mine association rules. This paper will present an efficient ARM algorithm by proposing a new technique to generate association rules from a huge set of items, which depends on the concepts of clustering and graph data structure, this new algorithm will be named clustering and graph-based rule mining (CGAR). The CGAR method is to create a cluster table by scanning the database only once, and then clustering the transactions into clusters according to their length. The frequent 1-itemsets will be extracted directly by scanning the cluster. Table ≥ 2, we build directed graphs for each cluster in the case of very huge amount of transactions.

Keywords: Apriori, CGAR, CD Algorithm, CBAR

PAPER-QR CODE

Corresponding Author: MR. AKHILESH B. BHUYAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Akhilesh Bhuyar, IJPRET, 2014; Volume 2 (9): 1025-1030

INTRODUCTION

Data mining is a tool that supports research and allows new assertions to be made by disclosing previously undisclosed details in Large amounts of data [11]. One of the most challenges in database mining is developing fast and efficient algorithms that can deal with large volume of data because most mining algorithms perform computation over the entire database and mostly the databases are very large.

A large item set is a set of items which appear often enough within the same transactions. In this paper, we introduce an algorithm called CGAR, which is fundamentally different from all the previous algorithms in the following points:

i. It reads the database of transaction only once to generate frequent 1-itemsets.

ii. It is scalable with all types of databases regardless to their size.

iii. It is easy to implement as it uses simple cluster table and a robust graph data structure. Due to growth of the data volume in the last decade, a set of different techniques for deletion of repetitive data and conversion of data to more usable forms has been proposed under the name of Data Mining.

In this paper we envision a distributed clustering algorithm which is scalable and provides cooperation while preserving a high degree of independency for each site. Clustering is a discipline aimed at revealing groups_ or clusters_ of similar entities in data. As clustering is an essential technique for data mining, distributed clustering algorithms were developed as part of the distributed data mining research Clustering or unsupervised learning, is the task of grouping together related data objects. Unlike supervised learning, there isn't a predefined set of discrete classes to assign the objects to. Instead, new classes, in this case called clusters, have to be found. There are a lot of possible definitions for what a cluster is, but most of them are based on two properties: objects in the same cluster should be related to each other, while objects in different clusters should be different. Clustering is a very natural way for humans to discover new patterns, having been studied since the ancient times. If it is regarded as unsupervised learning, clustering is one of the basic tasks of machine learning, but it is used in a whole range of other domains: The focus of this survey is the application of clustering algorithms in data mining.

Literature Review & Related work:

Association Rule Problem

Association rules, first introduced in 1993, are used to identify relationships among a set of items in a database, it was used in the sale transaction databases domain, and so there should be a set of [m] distinct items I = {I1, I2, …, Im}, and a database of transactions D, where each transaction T has a unique identifier TID, and contains a set of items such that I. An association rule is an implication of the form X Y, where X and Y are subsets of I, and they are disjoint, that is, X∩Y = X and Y are sets of items called item sets. The rule X Y holds in the database D with confidence c, if c% of transactions in D that contain X also contain Y. The rule X Y has support s in the transaction set D, if s% of transactions in D Contain X Y. Given the database D, in data mining, Apriori [is a classic algorithm for learning association rules.

This association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to

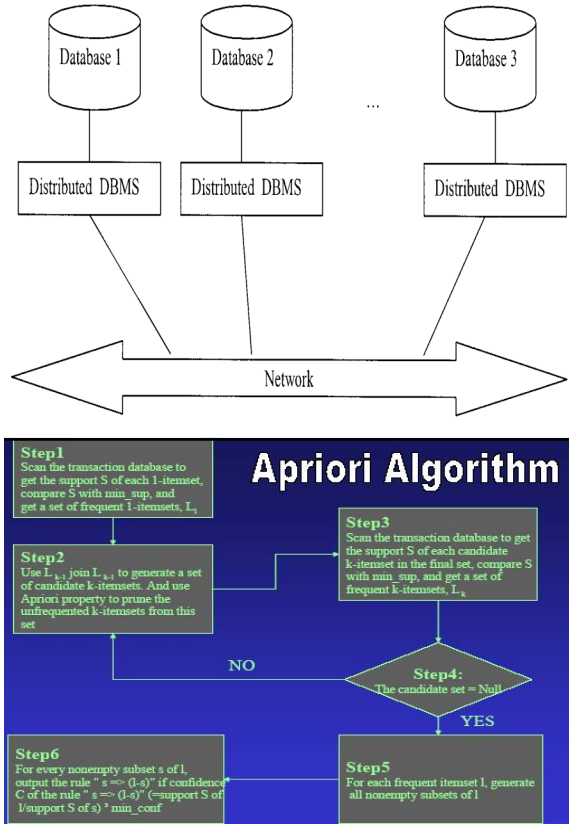

Fig 1: distributed Data Environment

A distributed database is a database in which storage devices are not all attached to a common CPU. It may be stored in multiple computers located in the same physical location, or may be

dispersed over a network of interconnected computer Collections of data (e.g. in a database) can be distributed across multiple physical locations. A distributed database can reside on network servers on the Internet, on corporate intranets or extranets, or on other company networks. The replication and distribution of databases improves database performance at end-user worksites. To ensure that the distributive databases are up to date and current, there are two processes: replication and duplication. Replication involves using specialized software that looks for changes in the distributive database. Once the changes have been identified, the replication process makes all the databases look the same. The replication process can be very complex and time consuming depending on the size and number of the distributive databases. This process can also require a lot of time and computer resources. Duplication on the other hand is not as complicated. It basically identifies one database as a master and then duplicates that database. The growth of the data volume in the last decade was so huge that it in a way hindered the attempts involving the efficacious understanding and use of data. Different techniques for knowledge inference, deletion of repetitive data and conversion of data to more usable forms has been proposed under the name of Data Mining. To name decision trees, associative rules, Bayes Networks and clustering.

Analysis of Problem:

Previous studies in data mining have presented efficient algorithms for discovering association rules. But the main problem in the first algorithms is the need to do multiple passes over the datasets to generate frequent item sets. The Apriori association rule algorithm [2] can discover meaningful item sets and build association rules within large databases, but a large number of the candidate item sets are generated from single item sets and this method also needs to perform contrasts Different strategies were developed after that to improve the process of generation association rules, as in FPGrowth [8], which outperforms all candidate-set-generation and- test algorithms as it mines frequent patterns without candidate generation, but it still have problems in the case of no common prefixes within the data items. Another technique is the sampling algorithm which reduces the number of database scans to a single scan, but still wastes considerable time on candidate item sets [5]. A third algorithm is the dynamic item set count (DIC) algorithm [6] for finding large item sets, which uses fewer passes over the data than classic algorithms, and yet uses fewer candidate item sets than methods based on sampling [5]. In addition, the column-wise apriori algorithm [10] and the tree-based association rule algorithm [4], transformed the storage structure of the data, to reduce the time needed for database scans, improving overall efficiency. Finally, the partition algorithm infrequent candidate item sets. Pork et al. proposed an effective algorithm DHP (direct hashing

and pruning) [3] for the initial candidate set generation. This method efficiently controls the number of candidate 2-itemsets, pruning the size of database[8].

Proposed Work and Objectives:

Cluster and graph based association rule

Although, the Cluster-based Association Rule (CBAR) algorithm [1] outperforms Apriori algorithm as it scans the database only once, but the opportunity to enhance cluster based algorithms still available by providing an efficient graph data structure to simplify the process of generating frequent k item sets, where k ≥ 2. In this paper, we present a new algorithm called clustering and graph-based association rule (CGAR), for efficient association rules mining, which

Overcome the drawbacks of the previous algorithms. The items should be given sequential numbers to simplify the process of building the graph; this must be taken in consideration as an important action before applying our proposed algorithm. CGAR scans the database of transactions only once to build the clustering table as a two-dimensional array where the columns represent items and the rows represent transactions' IDs (TIDs). The contents of the table consist of 0 or 1 to indicate the absence or presence of an item in a transaction, as the graph is completed, the set of frequent 2-itemsets are generated, and it will be direct from the graph traversing to generate frequent k-item sets, such as k ≥ 3. CGAR will deal with only one type of ARs, that is, BooleanARs.

Application

Association rule apply:

- Supermarket
- Purchases made using credit or debit cards
- Banking service ,unusual combination of insurance claims
- Medical patient histories.

CONCLUSION:

The project "Performance Improvement in using Clustering in Distributed Database" aims at accomplishing the task of allowing the project manager to maintain the project details. It also helps in maintaining the time details of each project. The system provides a graphical user interface, which helps all the employees to know he project details. It also generates reports, which gives detailed information about the clients of the company, senior developer, junior developer. Their size including the team leaders, tester, project developer etc.

REFERENCES:

1. Human Wu, Zhigang Lu, Lin Pan, Rongsheng An Improved Apriori-based Algorithm for Association Rules Mining Sixth International Conference Knowledge Discovery 2009.

2. Yuh-Jiuan Tsay, Jiunn-Yann Chiang, CBAR: an efficient method for mining association rules, Knowledge-Based Systems 18 (2005) 99–105.

3. R. Agrawal, T. Imilienski, A. Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993 pp. 207–216.

4. S. Ayse Ozel and H. Altay Güvenir, An Algorithm for Mining Association Rules Using Perfect Hashing and Database Pruning, (2000).

5. R. Agrawal, R. Srikant, Mining sequential patterns, Proceedings of the 11th International Conference on Data Engineering (ICDE), 1995.

6. F. Berzal, J.C. Cubero, N. Marin, J.M. Serrano, TBAR: an efficient method for association rule mining in relational databases, Elserier Data and Knowledge, Engineering 37 (2001) 47–64.

7. S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, ACM SIGMOD Conference on Management of Data, Tuscon, Arizona, 1997 pp. 265–276.

8. Ashok Savasere, Edward Omiecinski, and Shamkant Navathe. An Efficient Algorithm for Mining Association Rules in large databases. 1995

9. Han, J., Pei, J., Yin, Y: Mining frequent Patterns without Candidate Generation. In: ACM-SIGMOD, Dallas (2000)

10. Show-Jane Yen And Arbee L.P. Chen, A Graph-Based Approach For Discovering Various Types Of Association Rules, IEEE Transactions On Knowledge And Data Engineering, Vol. 13, No. 5, September/October 2001.