Algorithm for Distributed Database using Association Rule Mining

Prof. Akhil Anjikar[1] Samiksha Patankar[2] Seema Belgaonkar[3] Romy Kamble[4] Palash Nashine[5]
[1]Assistant Professor [2,3,4,5]Student
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]Rajiv Gandhi College of Engineering & Research, Nagpur, Maharashtra, India

Abstract— In current networking and communication environment maintenance of huge data is big concern. Application requiring huge data processing which leads towards several problems such as handling of huge amount of data monitoring of data, processing time for a database and growing databases. In distributed databases as amount of data is very large it effects the pre-processing of the system. In proposed system an algorithm is design to raise the pre-processing of a resulted data over distributed database. In this processing is done on large quantity of a data and collects the processed data for future. The optimize resulted database can be used for quick searches. Finest possible solutions are obtained for Distributed Databases.

Key words: Data Mining, Association Rule, Database Management, Apriori Algorithm

I. INTRODUCTION

Data mining technology aim to find useful patterns from large amount of data. Data mining is the process of analysing data from different angles & getting useful information about data. The data mining can help in predicting a trend or values, classifying, categorizing the data & in finding correlations, patterns from the dataset. The overall goal of data mining process is to extract information from a dataset & transform it into an understandable structure for future use.

The goal of the Data mining process is to extract information from a dataset & transform it into an understandable structure.

Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold [2].

Association rule mining consists of two stages:

1) The discovery of frequent item sets.
2) The generation of association rules.

It follows, that in the vast majority of cases, the discovery of the frequent set dominates the performance of the whole process. Therefore, we explicitly focus the paper on the discovery of such set.

Need for development of Distributed system for mining of association rules because of its unique properties:

1) Databases or data warehouses may store a huge amount of data. Mining association rules in such databases may require substantial processing power, and distributed system is a possible solution.
2) Many large databases are distributed in nature. For example, the huge numbers of transaction records of hundreds of Sears's department stores are likely to be stored at different sites.

II. RELATED WORK

Association Rule Mining (ARM) has been the area of interest for many researchers for a long time and continues to be the same. It is one of the important tasks of data mining. It aims at discovering relationships among various items in the database. Association rule mining is the concept to find frequent patterns, interesting correlations, and associations among sets of items in the transaction databases or other data repositories [1]. Given a set of transactions, association rule mining aims to find the rules which enable to predict the occurrence of a specific item based on the occurrences of the other items in the transaction [1].

Due to the rapid evolution of data collection and storage technologies, extracting knowledge and hidden patterns from stored data has become a major necessity for individuals, companies, and government agencies. However, applying data mining techniques to extract information is considered a challenge when the data is distributed over multiple owners, and each data owner is concerned about the privacy of individuals in his data. Privacy-Preserving Data Mining (PPDM) techniques has been utilized in the context of distributed computing to protect the confidentiality of the data of each provider, while still enabling the providers to perform data mining tasks, such as frequent itemsets mining and association rules mining, on the distributed data[2].

Association rule mining is a one of the most important technique in data mining. It extracts significant patterns from transaction databases and generates rules used in many decision support application. Modern organizations are geographically distributed. Using the traditional centralized association rule mining to discover useful patterns in such distributed system is not always feasible because merging data sets from different sites into a centralized site incurs huge network communication and time costs. Optimized Distributed Association Rule Mining (D-ARM) based on vertical partitioning. The existing D-ARM algorithms have lots of communication overhead, which is a major issue for concerning. The proposed approach minimizes this communication overhead and it is based on total count [3].

An Efficient Approach of Association Rule Mining on Distributed Database Algorithm applications requiring huge data processing have two main problems, one a massive storage and its supervision and next processing time, when the quantity of data increases. Distributed databases determine the first trouble to a huge amount but second problem increase. Since, current stage is of networking and communication and community are involved in maintenance huge data on networks, therefore, proposed an algorithm to raise the throughput of resulted data over distributed databases [4].

Algorithms for mining association rules from relational data have been well developed. Several query

languages have been proposed, to assist association rule mining such as the topic of mining XML data has received little attention, as the data mining community has focused on the development of techniques for extracting common structure from heterogeneous XML data. For instance, proposed an algorithm to construct a frequent tree by finding common subtrees embedded in the heterogeneous XML data. On the other hand, some researchers focus on developing a standard model to represent the knowledge extracted from the data using XML [5].

III. PROPOSED WORK

To minimize processing time of huge databases.

We need a mechanism which can process large quantity of the data from distributed resources and collect the process data for user in resulted database [1].

The optimize resulted database can be used for quick searches. At the end finest possible solution are obtained for distributed databases.

A. Algorithm

Step 1:
Generation or collection databases
Ex:
Division:
1) Turnover manufacturing
2) Sales by product
3) Total sale
4) product sales

Step 2:
Using Association Rule Mining and support count of the data we will generate frequency item sets.
Ex:
1) No of division contributing to the maximum turn over.
2) No of manufacturing industries for food product with maximum sale.
3) Which basic metal is having maximum sale.

Step 3:
Generate Association Rule from. Step 2
Ex:
1) Maximum turnover from exporting food products from division 10.

Step 4:
Store rule in resulted database.
Step 5:
Implementation
1) Generate query from data consumer.
2) Search the items accordingly to the attributes mentioned in consumer query in resulted database.

If
Found data provider will return the information to the data consumer
Else
Follow step 2
Follow step 3
3) Return result to data consumer

B. Diagram

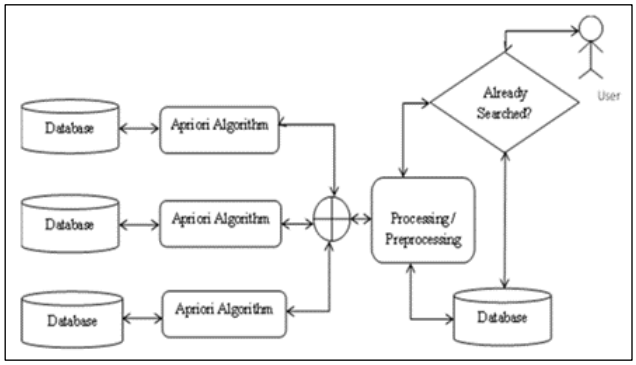

Fig. 1: Methodology

1) Association Rule Mining:

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Given a set of transactions, association rule mining aims to find the rules which enable to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

Association rule mining is the data mining process of finding the rules that may govern associations and causal objects between sets of items. So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. For example, peanut butter and jelly are often bought together because a lot of people like to make PB&J sandwiches.

2) Apiori Algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional database. It proceed by identify the recurring individual items in the database as well as extending them to bigger itemsets as long as goes item set come out adequately often in the database. The frequent item set examine by aprori and can be used to create a conclusion association rules which depict interest to general trends in the database. Apriori is design to work on databases containing transaction (for example collections of object bought by customers).other algorithms are plan for ruling association rule in data having no transactions or having no time stamp. Each transaction is seen as set of item. Apriori uses a "bottom up" method, where humorous subsets are extensive one item at an instance and groups of candidates are experienced alongside the data.

3) Distributed Database:

A distributed database is a database in which storage devices are not all attached to a common processor. It may be stored in multiple computers, located in the same physical location or may be dispersed over a network of interconnected computers. A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network. In a distributed database, there are a number of databases that may be geographically distributed all over the world. A distributed DBMS manages the distributed database in a manner so that it appears as one single database to users.

4) Preprocessing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-

world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications.

5) Processing:

Distributed processing is a phrase used to refer to a variety of computer systems that use more than one computer (or processor) to run an application. This includes parallel processing in which a single computer uses more than one CPU to execute programs.

Distributed processing refers to local-area networks (LANs) designed so that a single program can run simultaneously at various sites. Most distributed processing systems contain sophisticated software that detects idle CPUs on the network and parcels out programs to utilize them.

Another form of distributed processing involves distributed databases. This is databases in which the data is stored across two or more computer systems. The database system keeps track of where the data is so that the distributed nature of the database is not apparent to users.

C. Performance Evaluation

1) Home Page:

In fig 2 we have created a form in visual studio in which there are seven tabs. The first tab on the extreme left hand is for browsing and retrieving the Dataset stored in excel sheet. After clicking on the browse data tab all the data is loaded.

Below it there is a second tab of frequent itemset generation which shows the contribution towards turnover of maximum sale and minimum sale .which industry is contributing more and which industry contribution is less. It is applicable where sell and buy is used. The results will be shown respectively in the maximum and minimum contributor tabs.

On the extreme right hand there is a tab for search query. In which if structured query is entered related to datasets it will show all the details of it.

There is a exit tab used for getting out of the form that has been created.

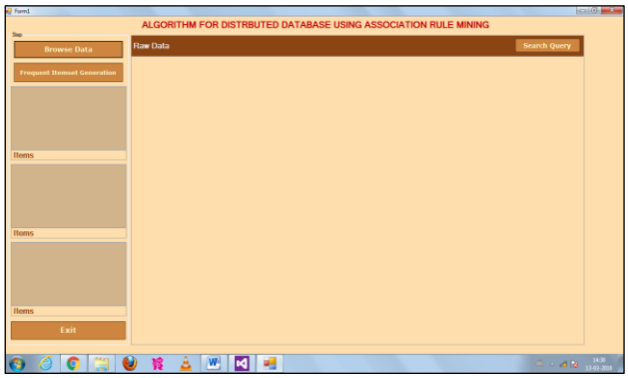

Fig. 2: Home Page

2) Prodcom:

In Fig 3 Prodcom after clicking on the browse data tab it helps us to achieve data connectivity which can be used for further processing.

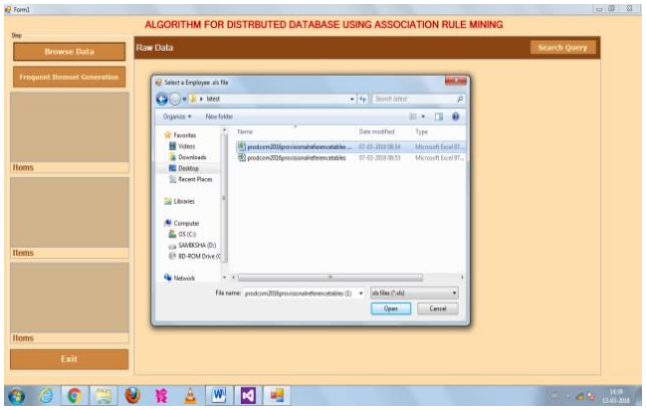

Fig. 3: Prodcom

3) Frequent Itemset Generation:

Fig.4frequent item set generation explains the contribution towards turnover of maximum sale and minimum sale .In this contribution turnover with maximum, minimum and moderate sale is shown. It gives an idea which industry is contributing maximum and mean. It is applicable where sell and buy is used.

The maximum and minimum turnovers are found by using association rule mining.

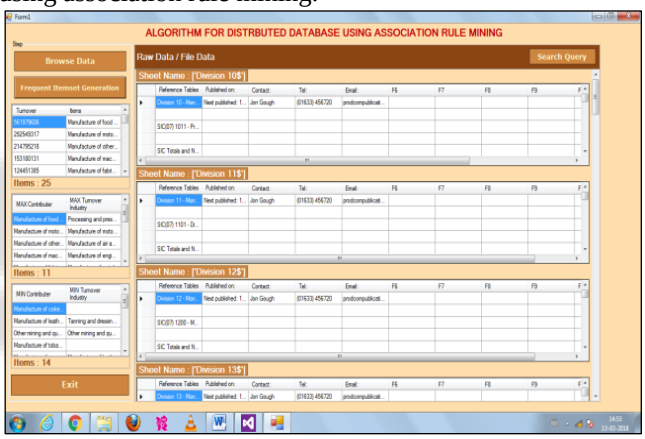

Fig. 4: Frequent itemset generation

4) Industry Wise Product Sale:

There are 11 industries that are contributing to maximum turnover .After left clicking on any one of the particular industry for example manufacture of food products a new window will open which shows the turnover of all the subdivisions.

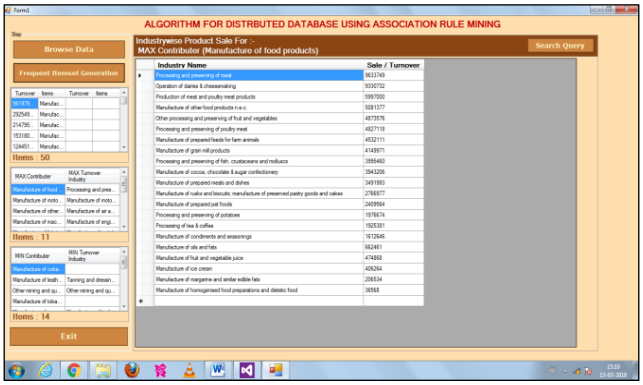

Fig. 5: Industry wise product sale

5) Search Query:

In the huge database it is difficult to see each and every information and sort out the information desired and required

by the user. So to minimize the time and access the required information a search query tab has been provided in which the user types the structured query information he/she wants and then it is processed and the results are shown. If the information entered in the query does not match with the datasets then it displays the result as no records shown. In other words the processing time is reduced from huge database and the information that is required by the user is shown ASAP.

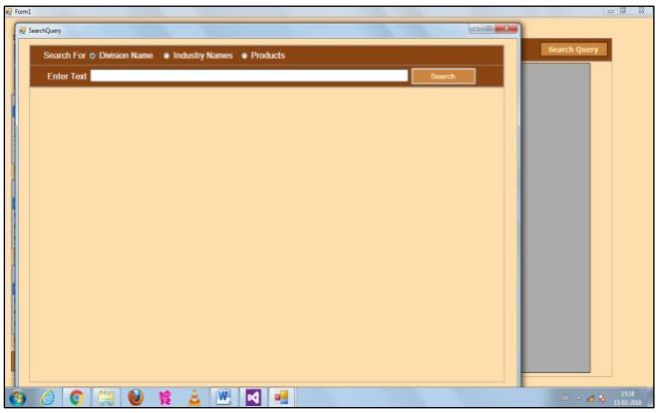

Fig. 6: Search query

6) Searching Information by Division Name:

There are 33 division in our data sets and if the user want to see the information as per the division (division means the turnover information of a particular domain) it will show the result which are domain specific.

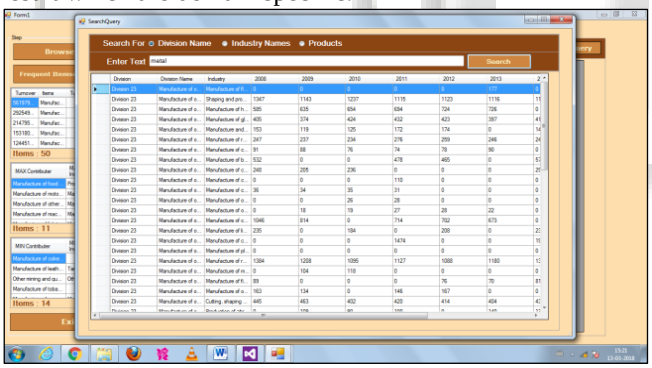

Fig. 7: Searching Information by Division Name

7) Searching Information by Industry Names:

There are 33 divisions in our data sets and if the user wants to see the information as per the particular industry it will show the detail result of its sub divisions.

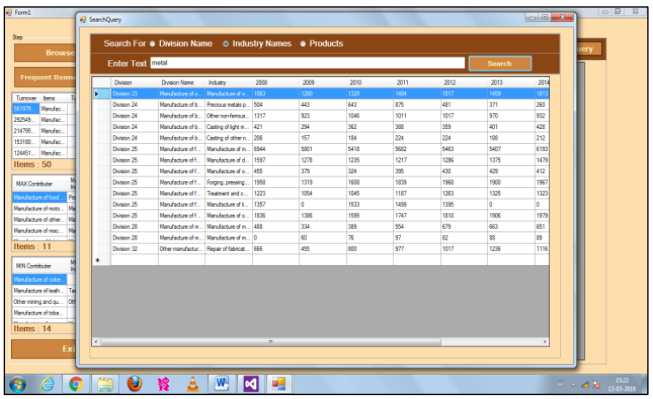

Fig. 8: Search Information by Industry names

8) Searching Information by Products:

There are 33 divisions in our data sets and if the user wants to see the information of the particular commodity by clicking on the product the filters will be applied and the query typed in the search bar will became specific to individual product irrespective of the industries.

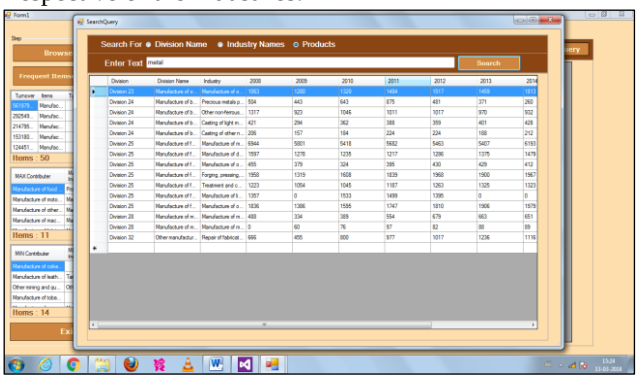

Fig. 9: Searching Information by Products

IV. CONCLUSION

The approach proposed in our paper uses an efficient method using association rule mining in distributed environment which reduces communication over head over the databases/ datasets.

In this approach the method of vertical portioning of dataset is used due to which communication between n datasets become easy. The result shows the efficiency to draw the conclusions by using association rules over large databases.

REFERENCE

[1] Gurneet Kaur et al, "Association Rule Mining: A Survey",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2320-2324

[2] Omar Abdel Wahab, Moulay Omar Hachami, Arslan Zaffari, Mery Gaby G. DagherVivas, "DARM: A Privacy-preserving Approach for Distributed Association Rules Mining on Horizontally-partitioned Data",ReseachGate, Conference Paper • July 2014

[3] Monika et al. "Optimization of Distributed Association Rule Mining Approach Based On Vertical Partitioning", International Journal of Computer Science Engineering (IJCSE)

[4] Neha Saxena, Rakhi Arora, Ranjana Sikarwar, Pradeep Yadav, "An Efficient Approach of Association Rule Mining on Distributed Database Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 81 – No.3, November 2013

[5] Dr(Mrs).Sujni Paul, "An Optimized Distributed Association Rule Mining Algorithm In Parallel And Distributed Data Mining With Xml Data For Improved Response Time.", International Journal of Computer Science and Information Technology, Volume 2, Number 2, April 2010

