

# **An Efficient Approach of Association Rule Mining on Distributed Database Algorithm**

**Neha Saxena**  
Asst. Professor  
CS/IT Dept  
IITM Gwalior

**Rakhi Arora**  
Asst. Professor  
CS/IT Dept  
IITM Gwalior

**Ranjana Sikarwar**  
Asst. Professor  
CS/IT Dept  
IITM Gwalior

**Pradeep Yadav**  
Asst. Professor  
CS/IT Dept  
IITM Gwalior

## **ABSTRACT**

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, researchers are suggest a range of novel algorithms to raise the throughput of resulted data over distributed databases. Within our research, we are proposing an novel algorithm to process large quantity of data at the a variety of servers and collect the processed data on customer machine as much as necessary.

## **Keywords**

Apriori algorithm, Association rules, parallel and distributed data mining.

## **1. INTRODUCTION**

Association rule mining is one of the mainly essential and fine researched methods of data mining. It aims to extort exciting correlations, common patterns, associations or informal structures amongst sets of objects in the transaction databases or additional data repositories. Association rules are broadly used in a range of areas such as telecommunication networks, market and hazard managing, inventory control etc [1]. Different association mining methods and algorithms will be momentarily introduced and compared afterwards. Association rule mining is to locate out association rules that suit the predefined least amount support and confidence from a database [3]. The trouble is decomposed into two sub problems. One is to discover those item sets whose occurrences go above a predefined threshold in the database; those item sets are known as frequent or large item sets. The second dilemma is to produce association rules from those large item sets with the constraints of negligible confidence [2].

The two most important approach for utilizing multiple Processors that have emerged in distributed memory is that each processor have a small private memory; [6]& public memory contained by the all processors right to use common memory. common memory structural drawing has many

accepted property. all processor has a direct and equivalent contact to all memory in the proposal.[4] In distributed memory structural plan every one processor possess local memory that can merely be access openly by that processor.

A Parallel purpose could be separated into numeral of subtasks and executed parallelism on separate processors in the system .although the appearance of a parallel application on a distributed system is typically focus on the allocation of the responsibilities comprising the application on top of the accessible processors in the scheme.[5] Association rule mining model amongst data mining numerous models, including Association rules, clustering and categorization models, is the mostly applied method. The Apriori algorithm is the mainly representative algorithm for association rule mining. It consists of plenty of modified algorithms that focus on civilizing its efficiency and accuracy.

## **2. LITERATURE REVIEW**

Association Rule Learning is a general technique used to discover associations amongst numerous variables. It is often used by grocery stores, retailers, and anyone with a bulky transactional database.[7] Association rules are then statements that assist out to discover associations between apparently isolated data in a relational database or supplementary data warehouse. An example of an association rule would be "If a customer buys a dozen breads, he is 80% likely to also purchase butter/jam." Association rules are produced by analyzing data for frequent then patterns and via the criteria sustain and confidence to identify the most important associations[9]. Support is an warning of how regularly the items come forward in the database.

In data mining, association rules are helpful for analyzing and predicting customer nature[8][9]. They play an significant role in shopping basket data analysis, item clustering, catalog design

Programmers use association rules to construct programs of machine learning[5]. Machine learning is a sort of artificial intelligence that seeks to assemble programs with the capability to develop into more competent without being explicitly programmed.

Algorithms for mining association rules from relational data have been developed. numerous query languages have been

planned, to assist association rule mining such as the issue of mining XML data has acknowledged very little concentration, as the data mining society has paying attention on the progress

The PADMA tool is an article analysis device executing on distributed environment, based on co-operative agent. It works without any relational database underside.

### 3 ASSOCIATION RULE MINING ALGORITHMS

An association rule implies definite association interaction among a set of objects in a database. An association rule is an appearance of the A B form, where A & B are data frequent items[10]. The insightful sense of such a regulation is that transactions of the database which control A be liable to enclose B. Association rule is one in many of the data mining procedure used to huge data out concealed information starting datasets that can be use by an organization decision maker to get better on the complete earnings.

#### 3.1 Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceed by identifying the recurring individual items in the database as well as extending them to bigger item sets as long as those item sets come out adequately often in the database. The frequent item sets examined by Apriori and can be used to create a conclusion association rules which depict interest to general trends in the database. Apriori is designed to work on databases containing transactions (for eg. collections of objects bought by customers.)([11][12]. Other algorithms are planned for ruling association rules in data having no transactions or having no timestamps. Each transaction is seen as a set of items.

Apriori uses a "bottom up" method, where numerous subsets are extensive one item at a instance and groups of candidates are experienced alongside the data. Pseudo code below demonstrates the process of frequent item set generation of the Apriori algorithm.

#### 3.2 Distributed/Parallel Algorithms

Databases may accumulate an enormous quantity of data to be mine. Mining distributed association rules within the such databases management system might involve significant processing power [13]. A likely pledge to this dilemma can be a distributed system. Moreover, lots of databases are dispersed which may accumulate it more probable to utilize distributed algorithms. most important thing of association rules is the totaling of the set of giant frequent item sets in the database. dispersed computing of bulky item sets encounters a numeral of original tribulations. individual might guess locally huge article sets naturally, but a locally large frequent item set may not be internationally large[7].

Since it is very costly to transmit the whole data set to other sites, one alternative is to put on air all the counts of all the item sets. However, a database may take hold of very big combinations of item sets, and it will engage momentary a huge amount of communication.

#### 3.3 Optimized Distributed Association Rule

Association rule mining is an dynamic data mining research area. However, most association rules mining algorithms provide a centralized atmosphere. In contrast to previous

of techniques for extracting common arrangement from varied XML data.

ARM algorithms, optimized distributed association rule is a distributed algorithm for physically and logically distributed data sets that reduces communication overheads[14][5]. Current organizations are physically distributed. usually, each site close by supplies its ever increasing amount of day-to-day data. Using federal data mining to find out useful patterns in such organizations ' data isn't always possible because integration of data sets from different sites into a centralized site incurs huge network communication overheads. Data from these organizations are not simply distributed over various locations but also vertically disjointed, making it difficult if not impossible to unite them in a central position mining algorithm. To defeat these problems, we don't produce candidate support counts from the raw data set after the first pass. This method reduces the normal transaction length

### 4. PROPOSED ALGORITHM

We will focal point on telling the experiments planned to estimate the performance of the projected Data Structure Mining algorithm. At this time, Association ruling acting an important role. The purchasing of individual product when an additional product is purchased represents an association rule.

This algorithm developed to present here the distributed data at a very quick rate to the users engage flow of processing of data the same as follows. Customer demands the data from the crossing point given. Data demanded is transferred to the proxy server, somewhere it is initially checked in the local database for simplicity of access, if the data is accessible, then provide to the user and occurrence of data is incremented, if not data is transfer to the various Distributed databases using multithreaded atmosphere. The variety of servers throw the number of winding up to the proxy server, where it is joint cooperatively to come across the infrequent item set for the searched charge Item customer /Proxy Server .

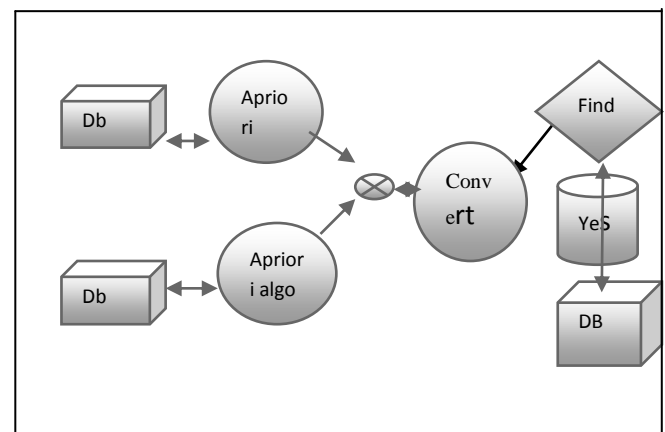


Fig 1: flow chart of DB algorithms.

### 5. DATA SET

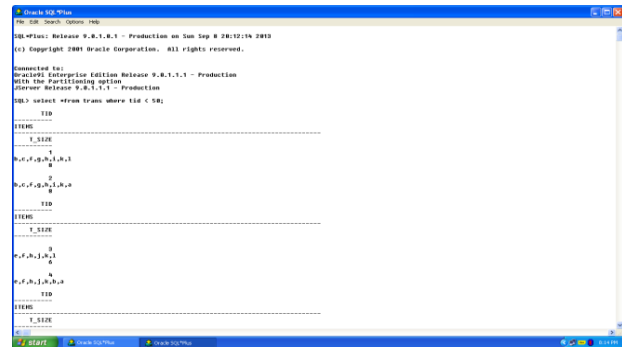
The experimentation is carried out with the help of synthetic datasets that are generated through the use of a dataset generator that is publicly available. A data set is a gathering of data, frequently presented in tabular shape. Each column represents a particular variable. Each row corresponds to a specified associate of the data set in question. It gives values to every variable, such as transaction id and transaction of an object. Each value is known as a element. The data set might consist of data for one or more members, equivalent to the

number of rows. . For example, consider a sample database as shown in Table1.1

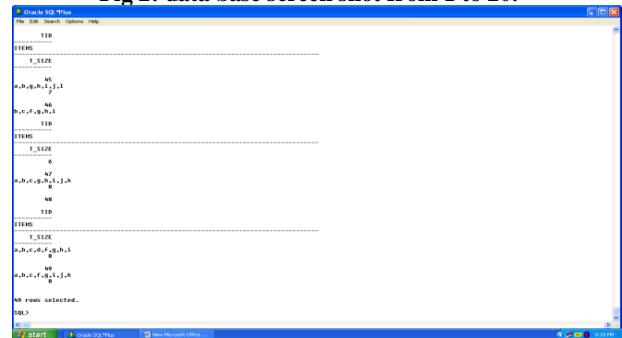
**Table 1: table of item number and item sets**

S.N	Item Number	Item Set
1	I1	a,b,c,d,e,f
2	I2	d,b,a,e,c,e
3	I3	b,g,i,h,j
4	I4	c,h,a,b,e
5	I5	a,b,c,e,i
6	I6	d,e,i,j,d
7	I7	e,g,h,i ,j
8	I8	a,d,c,j,i,b,e
9	I9	c,a,b,f,g,h
10	I10	a,e,i,f,k,l,b

The database created at the run time is shown below



**Fig 2: data base screen shot from 1 to 20.**



**Fig 3: data base screen shot from 20 to 50.**

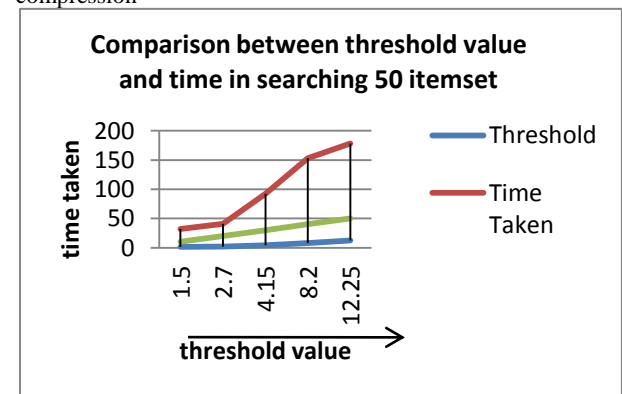
**Table 2: Data collected Threshold Vs Time**

S.N	Threshold	Time Taken	Item set
1	1.5	32	10
2	2.178	41	20
3	4.15	92	30
4	8.2	153	40
5	12.25	178	50

## 6. RESULTS & IMPLEMENTATION

The dilemma of mining association rules is to produce all policy that have support and self-confidence better than or equal to some customer specified lowest support and least confidence threshold correspondingly. We have evaluated the performance of our proposed algorithm (DB algorithm) by comparing its execution time with the threshold value of the existed algorithms.

Because of the huge size of data and quantity of working out involved in data mining, high-performance computing is an indispensable constituent for any successful large-scale data mining applications. We have applied our proposed algorithm on this database, following results have been come out then the result shown in fig the result between the threshold values and time taken in searching the 50 item set and show the compute value to table 1.1. Show the comparisons graph1.1 .or transaction record & time set in shown in table1.2 fig or total compression



result in as shown in fig

**Fig 4: Comparison between threshold value and time in searching 50 item set**

**Table 3: Data collected no of Records processed**

S.N	Number of records	Time taken
1	25	158
2	30	297
3	35	412
4	40	703
5	45	978

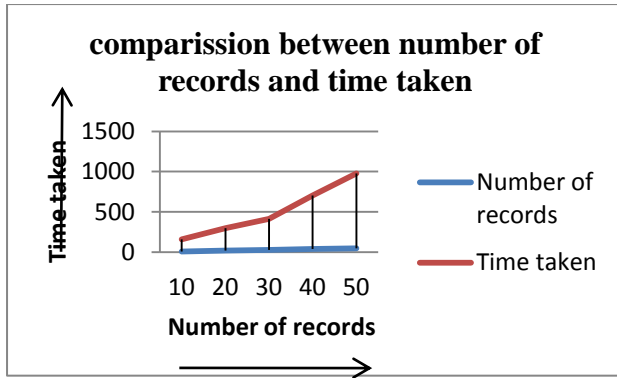


Fig 5: comparison between number of records and time taken

## 7. COMPARISION

Comparison between Apriori Algorithm and new proposed algorithm(DB algorithm) For the item sets given above,we have compared both the algorithm i.e apriori as well as new proposed algorithm DB algorithm on same item sets,and following result have come out. This is an example based on the following transactions in the database. First we are apply the apriori algorithm then distributed database algorithm(DB) to find searching item set . based data base architecture to find the frequent item sets. This proposed work highlights the important aspects of system implementation, including the technology choice, algorithm implementation and other interesting implementation solutions. The main objective of this stage is to transform the design solutions into working model. The comparison is shown in table 1.3 apriori and data base algorithms. And the comparison result shows that our purposed algorithm is better then apriori algorithms.

Table 4: comparisons between apriori and db algorithms.

S. N	Thres hold	Time Taken		No. Of Reco rds	Time Taken	
		APRI ORI	D B		APRI ORI	D B
1	1.5	41	32	25	167	158
2	2.178	53	41	30	302	297
3	4.15	107	92	35	423	412
4	8.2	164	153	40	721	703
5	12.25	189	178	45	981	978

The proposed algorithm is applied on 50 item sets and the time taken by DB algorithm is much less than apriori algorithm.

## 8. CONCLUSION & FUTURE ENHANCEMENTS

Association rule mining is a significant management. The Optimized Distributed Association Mining Algorithm is worn for the mining progression distributed background. The reply time throughout the announcement and calculation factors are measured to conquer the finer arrival time, batch of processors in a single environment. As the mining process is ended in parallel an finest possible solution is obtained. The various graphs show the processing time as estimated and generate the results as per the requirements of the users. Fast response time as shown in the graphs shows that the proposed algorithm generates the outcome as needed. The upcoming improvement of this is to work about on proxy server to permit users to access new data searched even when the data is found in the neighborhood. The exploitation of conventional approach will be hard to collect the latest demand for data mining, so the new data mining algorithm proposed in this paper is meaningful. This paper increases data mining helpfulness significantly. This DB method can solve the algorithm space problem in our environment. The response time with the statement and calculation factors are measured to attain an enhanced response time. The routine analysis is completed by rising the quantity of processors in a distributed environment. As the mining course is done in similar an optimal resolution is obtained. The prospect enrichment of this is to collect the same dataset and uncover out the facts extracted out of that. A visual analysis can also be made for the same.

## REFERENCES

- [1] Dr .Sujni Paul, Associate Professor, Department of Computer Science, Karana University, Coimbatore 645 , Tamil Nadu, India
- [2] R. Agrawal and R. Srikant , "Fast Algorithms for Mining Association Rules in huge amount of Database, Conf. on very Large Database system .
- [3] R.J Agrawal and J.C. Shafer , "Parallel Mining of association Rules in data mining , Distributed Systems held in March 2005
- [4] D.W.K Chung ,"Efficient Mining of Asso. Rules in Distributed DB, "IEEE Knowledge & Engg., vol. 7,
- [5] D.W.K Chung,"A Fast Dis Distributed Information computing system, IEEE CS Press, 1997,
- [6]Albert Y.N Zomaya, Tark El.J-Ghazawi, Ophir Frierer, 'Distributed Computing for Data Mining'iee conference, held in 1996. International Journal of Computer application and Information Technology, Volume 3, Number 3, April
- [7] A. Prodromidis, P. Chan, and S. Stolfo. Chapter Meta learning in Parellal distributed data mining systems: Issues and approaches. AAI/MIT Press, 2001.
- [8] Morgan Kaufmann, 1996, pp. 432 Proc. ACM SIGMOD 1-12. 2010- 99 Proc. 20th Int'l 16 IEEE tribute Proc. Parallel and 432-444. national conference of Computer system and application, Volume 4, Number 3, April 2010 .
- [9]M.J. Zaki et al., Parallel Data Mining for Association Rules in partial Memory. ,tech. report TR 618, Computer Science Dept., Univ. of Rochester, 1997

- [10] D.W. Cheung et al. "Efficient Mining of Association Rules in mass Database 'IE Knowledge information Eng., vol. 8, no. 6, 1996, pp.916-923;
- [11] A. S and R. Wolff , "Communication-Efficient Distributed Mining of Association Rules '. SIGMOD national Conf. on Mgmt.of Data, ACM Press, 2001, pp. 47-48.
- [12] T.K Imielinski and A.M Virmani. MSQL: A query language for database management mining. 1999.
- [13] H.Kargupta, I.Hamzaoglu, and Brian Stafford. Scalable, distributed data mining-agent architecture. In Heckerman et al. [8], page 21.
- [14] R. Meo, G.S Psaila, and S. K Ceri. A new SQL like operator for mining association rules. In The VLDB Journal, pages 156–161,
- [15] T. Shntani and Kitsuregawa , 'Hash-on Based Parallel Algorithms for Mining Association Rules mining'. Conf. Parallel and Distributed Systems, IEE Press, 1998. 18-34;
- [16] Huan, Zhing Lu, Rongsng Xu, WenbJiang, 'Apriori-based Algorithm for Association Rules Mining', 9th national Conference on Fuzzy Systems Knowledge Discovery, IEEE Society community, 2007
- [17] Rupali Haldukar 'Optimization of Association Rule Mining with Genetic Algorithm', International Journal of Computer Science and Engineering (IJCSE), Vol. 2, Issue. 1, May 2011
- [18] Huiying Xiawei Le, 'The Research on Improved Association Rules Mining Apriori Algorithm' 2012 sixth International Conference on Fuzzy Systems an & Knowledge Discovery
- [19] Mrs. R. Sumithra, Sujni Paul, 'Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge '2011 third International conference on Computing and Networking Technologies.