Comparative Study of Apriori and FP-Growth Algorithm in Horizontally Partitioned Data using Client Server Approach

Ankita Sahu[1], Priyanka Pitale[2]

[1]*M. Tech. Scholar, Department of Computer Science & Engineering, RSR RCET, Bhilai, Chhattisgarh*

[2]*Assistant Professor, Department of Computer Science & Engineering, RSR RCET, Bhilai, Chhattisgarh*

***Abstract*- Data mining is one of the important fields of computer science, which deals with mining of important information from the data. Association rule mining is one of the association rule mining technique is getting very popular now a days and is getting much more attention from the researchers. Mining data from distributed database is also one of the important aspect of data mining very data of one site is not known to the other site. This research paper focus on Comparative Study of Apriori and FP-Growth algorithm in Horizontally Partitioned Data using Client Server Approach and software is implemented in NETBEANS IDE 8.2. Concepts of Client Server technology is being used for implementation of distributed database.**

***Index Terms*- Data Mining, Distributed database, Association rule mining, frequent pattern.**

I. INTRODUCTION

Mining Association rule from distributed database is getting attention of researchers now a day. The transactional data may be scattered among various sites and a centralized system can access these transactional data for mining global association rules with minimum support and confidence value given by user. The distribution of data can be vertically, horizontally and hybrid. In horizontal partitioning each site contains a subset of records of the original relation R. where as in vertical partitioning, each site contains only a subset of the attributes of a relation R. This paper address association rule hiding in distributed environment (Horizontally partitioned).

Agrawal introduced association rule mining. Global association rules can be found without disclosing any sensitive information [1].We can find the global confidence of an association rule BC→D by knowing the local support of BC and BCD, and the size of each database. The individual transactions need not be shared. The Apriori algorithm on a single site can be extended to distributed sites. If a rule has support> k% globally, it must have support > k% on at least one of the individual sites.

Using this global rule can be found, but the individual parties mustreveal which rule it supports [2]. The formula for support and confidence can be viewed as figure 1.

$$support_{AB\Rightarrow C} = \frac{\sum_{i=1}^{sites} support\_count_{ABC}(i)}{\sum_{i=1}^{sites} database\_size(i)}$$

$$support_{AB} = \frac{\sum_{i=1}^{sites} support\_count_{AB}(i)}{\sum_{i=1}^{sites} database\_size(i)}$$

$$confidence_{AB\Rightarrow C} = \frac{support_{AB\Rightarrow C}}{support_{AB}}$$

Fig 1: Support and confidence calculation

II. BACKGROUND

This section contains the terminology used in this paper like mining of association rule, distributed mining of association rule.

A. Apriori Algorithm

Apriori is designed to work on databases containing transactions [3]. The purpose of the Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. Output of Apriori is sets of rules that tell us that how many items are contained in sets of data.

The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Number of transaction is present in each set of data. Initial scan/pass of algorithm counts occurrence of each item in order to determine the frequent items set. Next scan K consists two phases.

1)In first phase, Candidate item set CK is generated using frequent item set Lk-1 found in (K-1)th pass. This is candidate generation process in Apriori Algorithm.

2) In second phase database is scanned to find support for Candidates CK. In next step, it prunes the candidates which have an infrequent sub pattern and keep only subset of candidate sets which are already identified as frequent items sets. Output of Apriori algorithm generates sets of rules thatidentify how often items are brought together in single set.

B. FP Growth Tree

FP Growth is one of the basic algorithm use for generate association rules[4]. FP growth is an approach based on divide and conquers method. The main purpose of this technique is to produce frequent item sets by using the combination of data attributes. It basically works on to generate frequent item set without candidate set generation [5]. The below figure demonstrate the concept of FP-Growth tree [6].

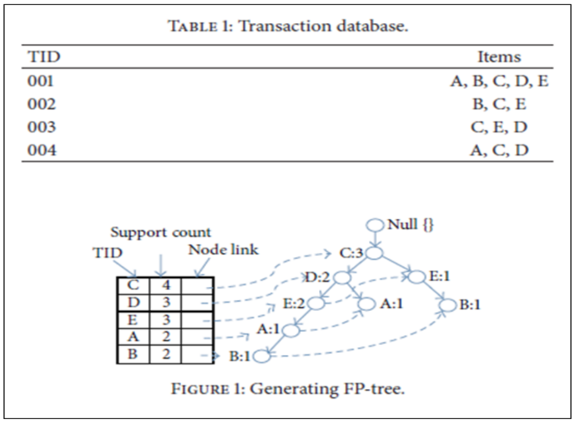


TABLE 1: Transaction database.

TID	Items
001	A, B, C, D, E
002	B, C, E
003	C, E, D
004	A, C, D

FIGURE 1: Generating FP-tree.

Fig 2: Example for demonstrating FP-Growth tree

C. Mining of Association Rule

The association rules mining problem can be defined as follows[7]: Let I = {i_1, i_2…I_n} be a set of items. Let DB be a set of transactions, where each transaction T is an itemset suchthat T⊆I. Given an itemset X⊆I, a transaction T contains X if and only if X⊆T. Anassociation rule is an implication of the form X ⇒Y where X ⊆I, Y⊆ I and X ∩ Y =ϕ. The rule X ⇒ Y has support s in the transaction database DB if s% of transactions in DBcontains X ∪ Y. The association rule holds in the transaction database DB with confidencec if c% of transactions in DB that contain X also contains Y. An itemset X with k itemscalled k-itemset. The problem of mining association rules is to and all rules whose supportand confidence are higher than certain user specified minimum support and confidence.

D. Distributed Mining of Association Rules

The above problem of mining association rules can be extended to distributed environments [8]. Let us assume that a transaction database DB is horizontally partitioned among n sites (namely S_1, S_2, S_n,) where DB = DB_1 ∪DB_2 ∪ …DB_n and DB_i resides at side S_i (1 ≤ i ≤ n). The itemset X has local support count of X.supi at site Si if X.supi of thetransactions contains X. The global support count of X is given as X.sup $=\sum_{i=0}^{n}$ X.supi. An itemset X is globally supported if X.sup $\geq(\sum_{i=0}^{n}|DBi|)$. Global confidence of a rule X ⇒Y can be given as {X ∪ Y}.sup/X.sup.

The set of large itemsets L(k) consists of all k-itemsets that are globally supported. The set of locally large itemsets $LL_i(k)$ consists of all k-itemsets supported locally at site S_i. GLi(k) = L(k) ∩$LL_i(k)$ is the set of globally large k-itemsets locally supported at site Si. The aim of distributed association rule mining is to find the sets L(k) for all k >1 and the support counts for these itemsets, and from this compute association rules with the specified minimum support and confidence.

Fast algorithms related for distributed association rule mining is given in Cheung et.al.[9]. their procedure for fast distributed mining of association rules (FDM) is summarized below.

(i) Candidate Sets Generation

Generate candidate sets CGi(k) based on GLi(k-1),itemsets that are supported by the Si at the (k-1)th iteration, using the classic apriori candidate generation algorithm. Each site generates candidates based on the intersection of globally large (k-1) itemsets and locally large (k-1) itemsets.

(ii) Local Pruning

(iii) For each X∈CGi(k), scan the database DBi at Si to compute X.supi. If X is locally large Si, it is included in the LLi(k) set. It is clear that if X is supported globally, it will be supported in one site.

(iv) Support Count Exchange
LLi(k) are broadcast, and each site computes the local support for the items in L(k)∪LLi(k).

(v) Broadcast Mining Results
Each site broadcasts the local support for itemsets. From this, each site is able to compute *L(k)*.

III. PROBLEM IDENTIFICATION

Now a day it is very important to mine important knowledge from the transactional data which are scattered among various sites like data of super market. This paper discusses the mining of association rule from different sites site$_1$, site$_2$ …site$_N$. Another problem is to identify the technology for maintaining the transactional data from different sites.

IV. PROBLEM SOLUTION

The solution of the problem is to use concept of client server technology for maintaining the transactional data from different sites and hence we can connect multiple clients with centralized server which calculate the frequent pattern using apriori and FP-growth algorithm and then association rule mining of global data.

V. IMPLEMENTATION OF PROPOSED SYSTEM

For implementation of the system we have used the concepts of client server technology through we can connect multiple client by specifying IP address and port number of server. This project can run of any network and many clients can connect at a time.

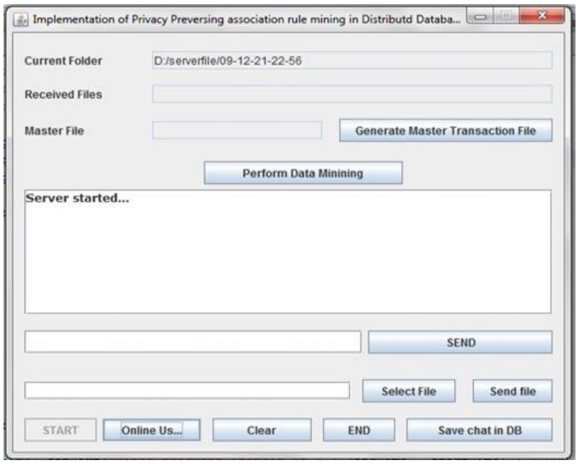
(i) Clients must have installed client program at their site and then clients can send their transactional data to the centralized sever for finding frequent itemset and association rule mining.

(ii) One data is received from all the sites then all the transactional data file can be merged horizontally to get master file.

(iii) After getting master file which contains transactional data from all the sites, we apply apriori and FP-Growth algorithms for finding frequent itemset and then we apply algorithm for finding association rules from frequent itemset.

VI. SNAPSHOTS

In order to established connect first we need to start server program so that it can accept connection request from multiple clients.



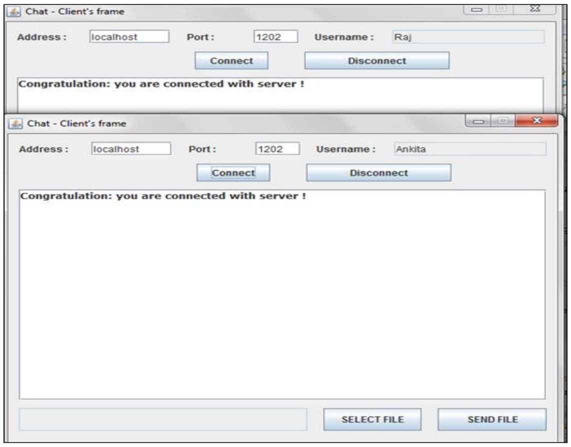
Implementation of Privacy Preversing association rule mining in Distributd Databa...

Current Folder: D:/serverfile/09-12-21-22-56

Server started...

Fig 3: Server form

The above figure shows that first we have to start server so that server can accept connection from the clients and as soon as server is started a default folder is created with name current system date and current time under server file folder in D drive.



Chat - Client's frame

Address : localhost Port : 1202 Username : Raj

Congratulation: you are connected with server !

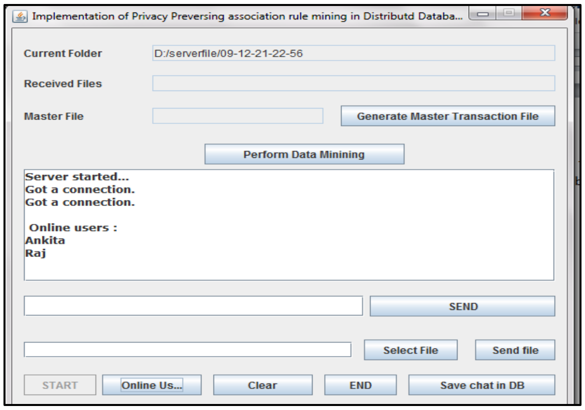
Chat - Client's frame

Address : localhost Port : 1202 Username : Ankita

Congratulation: you are connected with server !

Fig 4: Client form

Figure 4 shows that multiple clients can connect to server at a time by specifying server IP address, port number and client name like in figure two 2 clients Ankita and Raj is connected.


Implementation of Privacy Preversing association rule mining in Distributd Databa...

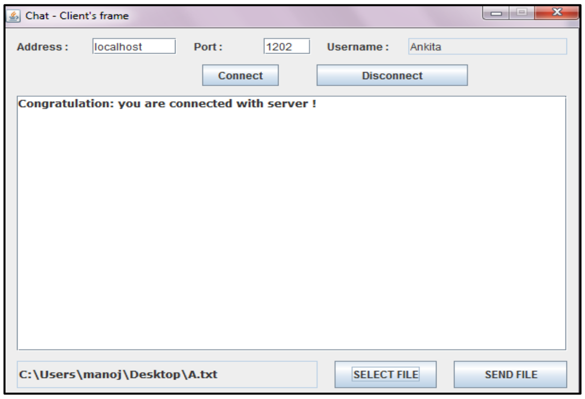
Current Folder	D:/serverfile/09-12-21-22-56
Received Files	
Master File	

```
Server started...
Got a connection.
Got a connection.

 Online users :
Ankita
Raj
```

Fig 5: Server form viewing clients connected

The above figure shows the details of clients connected with server as we click on online user button like here clients Ankita and Raj is connected.


Chat - Client's frame

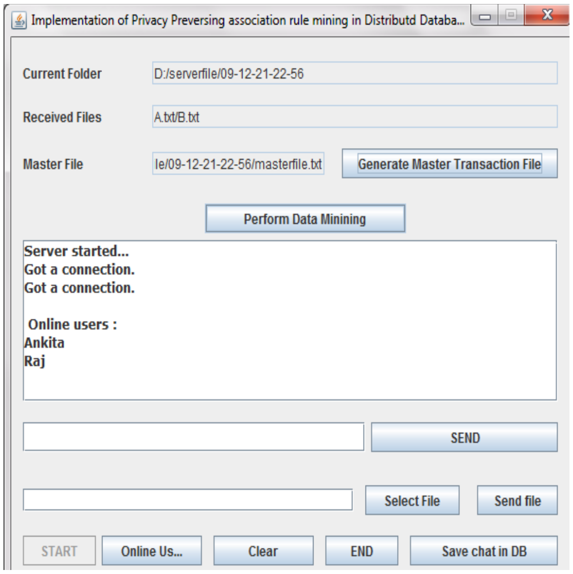
Address :	localhost	Port :	1202	Username :	Ankita

```
Congratulation: you are connected with server !
```

C:\Users\manoj\Desktop\A.txt

Fig 6: Clients sending transactional file to server

The above figure shows that client Ankitais sending transactional file A.txt to server by clicking SEND FILE button.


Implementation of Privacy Preversing association rule mining in Distributd Databa...

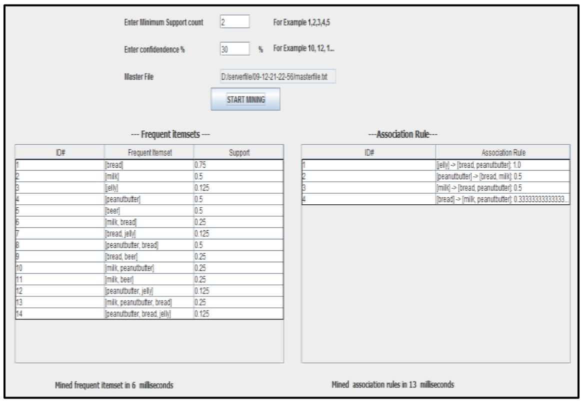
Current Folder	D:/serverfile/09-12-21-22-56
Received Files	A.txt/B.txt
Master File	le/09-12-21-22-56/masterfile.txt

```
Server started...
Got a connection.
Got a connection.

 Online users :
Ankita
Raj
```

Fig 7: Generating master transactional file from individual transaction file A.txt and B.txt

The above figure shows that master file is generated from given file A.txt and B.txt after merging files horizontally, now when we click on Perform Data Mining button form for performing data mining for opening association mining form.


Enter Minimum Support count	2		For Example 1,2,3,4,5
Enter confidendence %	30	%	For Example 10, 12, 1...
Master File	D:/serverfile/09-12-21-22-56/masterfile.txt		

--- Frequent itemsets ---

ID#	Frequent Itemset	Support
1	[bread]	0.75
2	[milk]	0.5
3	[jelly]	0.125
4	[peanutbutter]	0.5
5	[beer]	0.5
6	[milk, bread]	0.25
7	[bread, jelly]	0.125
8	[peanutbutter, bread]	0.5
9	[bread, beer]	0.25
10	[milk, peanutbutter]	0.25
11	[milk, beer]	0.25
12	[peanutbutter, jelly]	0.125
13	[milk, peanutbutter, bread]	0.25
14	[peanutbutter, bread, jelly]	0.125

---Association Rule---

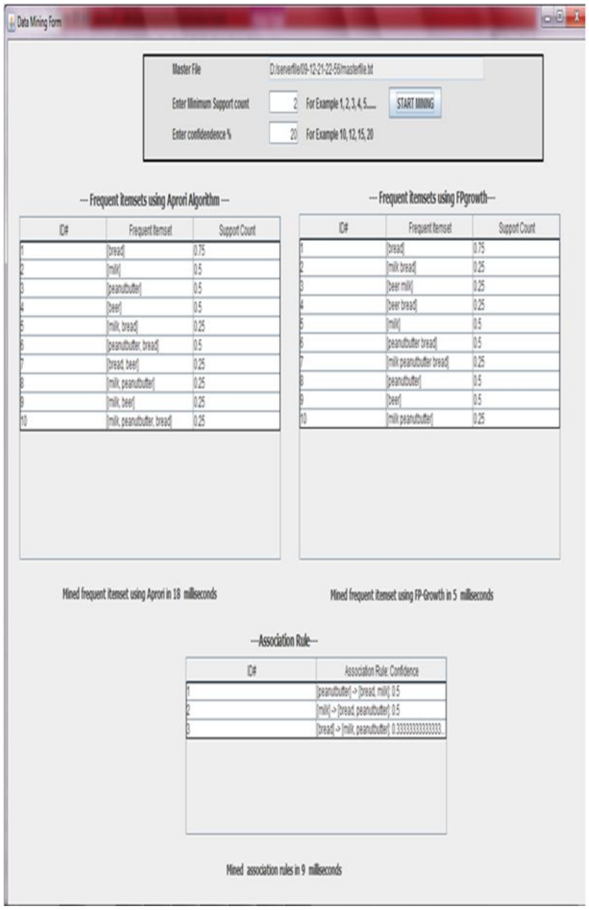
ID#	Association Rule
1	[jelly] -> [bread, peanutbutter]: 1.0
2	[peanutbutter] -> [bread, milk]: 0.5
3	[milk] -> [bread, peanutbutter]: 0.5
4	[bread] -> [milk, peanutbutter]: 0.33333333333333...

Mined frequent itemset in 6 milliseconds

Mined association rules in 13 milliseconds

Fig 8: Association Rule Mining form

Once Association rule mining form is opened, we put confidence and support value in the textbox after that we click on start mining button then frequent pattern and association rule is mined and shown in the table.


Data Mining Form

Master File	D:/serverfile/09-12-21-22-56/masterfile.txt	
Enter Minimum Support count	2	For Example 1,2,3,4,5....
Enter confidendence %	20	For Example 10, 12, 15, 20

--- Frequent Itemsets using Aprori Algorithm ---

ID#	Frequent Itemset	Support Count
1	[bread]	0.75
2	[milk]	0.5
3	[peanutbutter]	0.5
4	[beer]	0.5
5	[milk, bread]	0.25
6	[peanutbutter, bread]	0.5
7	[bread, beer]	0.25
8	[milk, peanutbutter]	0.25
9	[milk, beer]	0.25
10	[milk, peanutbutter, bread]	0.25

--- Frequent Itemsets using FPgrowth---

ID#	Frequent Itemset	Support Count
1	[bread]	0.75
2	[milk bread]	0.25
3	[beer milk]	0.25
4	[beer bread]	0.25
5	[milk]	0.5
6	[peanutbutter bread]	0.5
7	[milk peanutbutter bread]	0.25
8	[peanutbutter]	0.5
9	[beer]	0.5
10	[milk peanutbutter]	0.25

Mined frequent itemset using Aprori in 18 milliseconds

Mined frequent itemset using FP-Growth in 5 milliseconds

---Association Rule---

ID#	Association Rule: Confidence
1	[peanutbutter] -> [bread, milk]: 0.5
2	[milk] -> [bread, peanutbutter]: 0.5
3	[bread] -> [milk, peanutbutter]: 0.33333333333333...

Mined association rules in 9 milliseconds

Fig 9: Comparison of apriori and FP-Growth algorithm and association rule mining

VII. RESULTS AND COMPARISON

In the above diagram we have compared time taken by both apriori and FP-Growth algorithm for generation of frequent pattern on same transactional file.

Apriori algorithm takes more memory space as compared to FP-Growth algorithm and apriori algorithm scan DB many times whereas FP-Growth scans DB only twice hence it takes less time than Apriori algorithm.

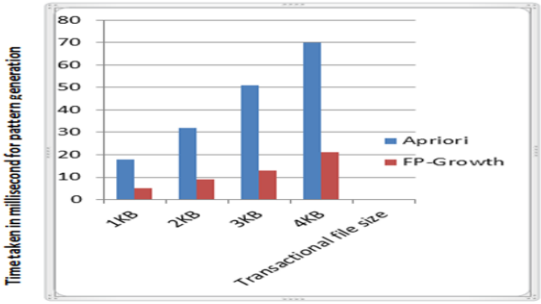


Fig 10: Comparison Chart of time and file size

The above chart showing comparison between apriori and FP-Growth algorithm in terms of size of transactional file and respective time taken by both algorithms (in terms of millisecond)

The above chart shows that the time taken by FP-Growth is more efficient in terms of time taken and it takes less time because the number of scan it takes is less than apriori algorithm.

VIII. CONCLUSION AND FUTURE WORK

In this paper we have presented a work on frequent pattern generation using apriori and FP-Growth algorithm and mining of association rule in distributed data base and the concept of horizontallypartitioning data has been used and also the concept of client server technology is being used. In future work on vertically partitioning can also be used.

REFERENCES

[1] Agrawal, Rakesh, and RamakrishnanSrikant. "Privacy-preserving data mining."ACMSigmod Record 29.2 (2000): pp439-450.

[2] Modak and ShaikhRizwana, "Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy", 7th International Conference on Communication, Computing and Virtualization 2016, pp993 – 1000.

[3] Sharda Darekar, Sandip Bankar, Prof. D. K. Chitre, "ASSOCIATION RULE MINING TO SECURE DATA IN DISTRIBUTED DATABASE", International Journal of Technical Research and Applications e-ISSN: 2320-8163, September 2015, pp 245-249.

[4] Abdullah Saad Almalaise Alghamdi, "Efficient Implementation of FP Growth Algorithm-Data Mining on Medical Data", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.12, December 2011, pp:7-16.

[5] Kuldeep Malik, Neeraj Raheja and Puneet Garg, Enhanced FP-Growth Algorithm, IJCEM Journal, April-2011.

[6] Yi Zeng, Shiqun Yin, Jiangyue Liu, and Miao Zhang, "Research of Improved FP-Growth Algorithm in Association Rules Mining", Hindawi Publishing Corporation Scientific Programming Volume 2015, Article ID 910281, 6 pages,January 2015.

[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", in Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile: VLDB,Sept. 12-15 1994. [Online]. Available: http://www.vldb.org/dblp/db/conf/vldb/vldb94-487.html.

[8] Murat Kantarciogluand Chris Clifton "Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data" IEEE Transactions on Knowledge and Data Engineering, January 2003.

[9] D. W.-L. Cheung, J. Han, V. Ng, A. W.-C. Fu, and Y. Fu, "A fast distributed algorithm for mining association rules," in Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA, Dec. 1996.

[10] Ankita Sahu, Priyanka Pitale,"Survey on Privacy Preserving Association Rule in Distributed Databases" in International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 3, March 17