ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Association Rules Mining in Vertically Distributed Databases

P.Kalaivani[1], D.Kerana Hanirex[2], Dr.K.P.Kaliyamurthie[3]

PG Scholar, Department of computer science, Bharath University, Chennai, India[1]

Assistant Professor, Department of Computer Science, Bharath University, Chennai, India[2]

HOD, Department of Computer Science, Bharath University, Chennai, India[3]

ABSTRACT: Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to make proactive, knowledge driven decisions. This process helps us to predict future trends and behaviours. Association Rule mining is one of the important and well researched techniques of data mining. We propose a protocol for secure mining of association rules in vertically distributed databases. The protocol [1] was experimented based on Fast Distributed Mining (FDM) algorithm and Secure Multi Party Algorithm over horizontally distributed databases. Our protocol is based on Apriori Algorithm and Secure Multiparty Algorithm simulated over vertically distributed databases. In this Paper, we have shown the experiment results of Apriori Algorithm and Secure Multi Party Algorithm over vertically distributed databases for finding Frequent Item Sets and efficiency over computation of arriving Association Rules.

KEYWORDS:Apriori Algorithm, Frequent Item Sets, Fast Distributed Algorithm (FDM), Association Rules, Multi Party Algorithm

I. INTRODUCTION

Here we study the problem of secure mining of association rules in vertically distributed databases. In such a setting, there are several databases where the transactional attributes are distributed across the databases. With the *vertical partitioning* approach, some of the columns of a relation R are apportioned into a base relation at one of the databases, and other columns are assigned into a base relation at another database. The relations at each of the sites must share a common domain so that the original table can be reconstructed. In “Market-Basket” example, one database may contain grocery items and other one may have clothing purchases. Using a key such as transaction date and transaction id, we can join these to identify relationships between purchases of clothing and groceries. Horizontal partitions support an organizational design in which functions are repeated, often on a regional basis, whereas vertical partitions are typically applied across organizational functions with reasonably separate data requirements.

Association rule is very important tool for mining process it has two special characteristics first one is support and another is confidence. Support gives total number of transactions of any particular item is occurring in datasets when confidence gives strength of a data in a dataset, support is probability of A and B and confidence is conditional probability.

The goal is to find all association rules globally with support s and confidence c for some given minimal support size s and confidence level c that hold in vertically distributed databases. There are two steps involved in mining of association rules. The first step is to detect those item sets whose occurrences exceed a minimum support threshold; those item sets are called frequent or large item sets. The second step is to identify association rules from those large item sets with the constraints of minimal confidence. As AIS algorithm [9] needs too many passes over the whole database, Apriori Algorithm for fast fetching of frequent Item sets.
The layout of the paper is as follows. In section 2, brief outlines of the papers referred to are given. Section 3 presents the Proposed System. In Section 4 Experiment Results are presented. In Section5, Conclusion is presented and lastly, Section6 arrives at the future enhancement

II. RELATED WORK

In [1], the author has presented a system for secure mining of association rules in horizontally distributed databases usingFast Distributed Mining (FDM) Algorithm and Secure MultiParty Algorithm. The Protocol facilitates enhanced privacy with respect to the protocol in [4]. In addition to that, it is simpler and is significantly more efficient in terms of communication cost, computational cost,and communication rounds. In [2], the authors have presented a protocol for discovering association rules between items in Large Databases. Experiment results have shown that Apriori Hybrid Algorithm is faster than AIS [7] and SETM [8] Algorithm [2]. In [3], the authors have offered FairplayMP, a generic system for Secure Multiparty Computation which is an extension of the Fairplay system that supported secure computation by two parties. In [4], the authors have presented a protocol for Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, which addresses the problem of computing association rules where the data may be distributed among various custodians, none of which are permitted to transfer their data to another site. Based on Fast Distributed Algorithm (FDM) and Secure Multiparty computation, Association Rules have been computed [4].In [5], the authors have proposed a system for privacy preserving Association Rule Mining in Vertically Partitioned Data. The system is demonstrated through a two-party algorithm for efficiently discovering frequent item sets with minimum support levels, without either site transmitting individual transaction values. In [6], the authors have done a comparative study for different vertical partitioning algorithm and shown how graph-based vertical partitioning algorithm has contributed towards the optimization of data fragmentation problem by providing an efficient way of improving performance of applications. In [7], the authors presented an efficient algorithm for extracting association rules between items in the databases. The Algorithm includes a novel estimation method, buffer management and pruning techniques. In [8], the authors show that at least some aspects of data mining can be carried out by using general query Languages such as SQL, PLSQL rather than by developing specialized algorithms [8]. In the earlier research work [9],[10],[11],[12],[13],[14],[15],[16] the author uses various techniques such as transaction reduction ,clustering and algorithms such as Apriori and fpgrowth for mining frequent item sets using different datasets were analysed and compared. In [17], the author has proposed the survey on Secure Mining of Association Rules in Vertically Distributed Databases.

III. PROPOSED SYSTEM

The proposed system uses two algorithms, namely Apriori and Multi Party for finding frequent item sets from vertically distributed databases. Association Rules are generated from the frequent items sets and grouped whose confidence is greater than the minimum threshold confidence called Strong Association Rules. The strong association rules are grouped in this way are displayed to the user.

The proposed system consists the following modules.

1) Product Details
2) Secure Mining of details
3) Combining databases
4) Frequent Item sets
5) Association Rules

1) Product Details
This module displays the product information maintained in the system. System maintains product id and its respective product name.

Fig.1 shows architecture of the proposed system.

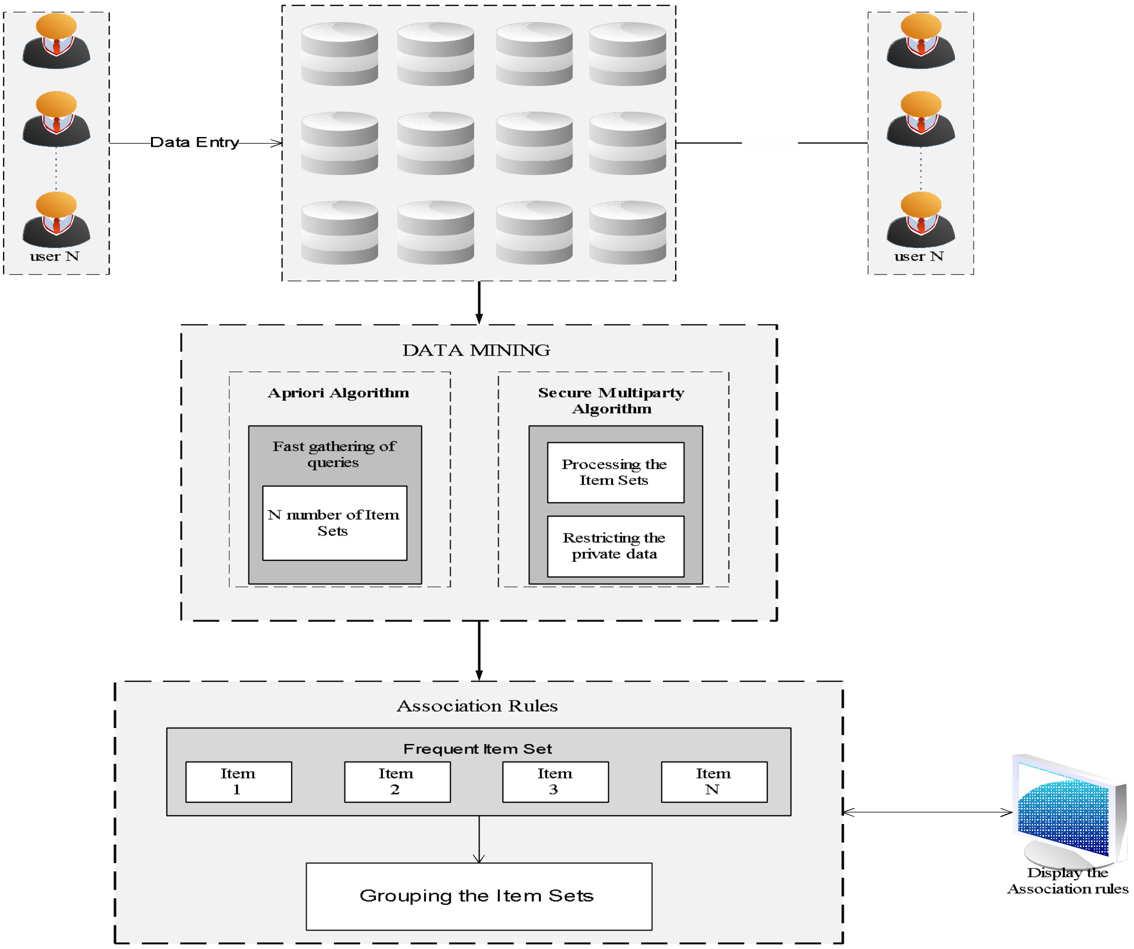


Fig. 1. System block diagram of Secure mining of association rules in vertically distributed enviroinment

2) Secure Mining of details
This module displays the transaction details maintained in the vertical databases without showing private attributes of the customer.

3) Combining Databases
This module does the process of combing the databases which are vertically distributed. Since transaction attributes are distributed across databases, the attributes are combined based on the transactions ids to identify Frequent Item Sets and Association Rules.

4) Frequent Item sets
This module displays the frequent item sets using apriori algorithm from the combined datasets. Frequent item sets are sets of items that have minimum support. Minimum Support is captured from the user and based on which, frequent Item sets are displayed to the user.

5) Association Rules
This module displays the Association Rules from the Frequent Item Sets found.
Steps for Association Rules:

- For each frequent itemset “l”, All nonempty subsets of l generated
- For every nonempty subset s of l, output the rule “s Implies (l-s)” if

support_count (l) / support_count(s) is greater than or equal to min_conf where min_conf is confidence threshold.

Apriori Algorithm
Apriori is a more efficient algorithm for association rule mining. Apriori algorithm was first presented by Agrawal in [Agrawal and Srikant 1994]. Apriori employs a different candidates generation method and a new pruning technique.

Key Concepts
- Frequent Item sets

All the sets which contain the item with the minimum support (denoted by$Li$for ith item set).

- Apriori Property

Any subset of frequent item set must be frequent.

- Join Operation

To find Large Item set Lk, a set of candidate k-item sets is generated by joining$Lk-1$ large item set with itself.

Algorithm Steps
Apriori algorithm having a two-step process.

a) The join step
To find Lk , a set of candidate k item sets is generated by joining large item set Lk-1 with itself. This set of candidate is denoted Ck.

b) The prune step
Ck is the superset of large item set Lk, that is, its members may or may not be repeated, but all of the frequent k-item sets are included in Ck. A scan of the databases to determine the count of each candidate in Ck would result in the determination of large item set Lk. (i.e. all candidates having a count not less than the minimum support count are frequent by definition, and therefore belongs to Lk)

Secure MultiParty Algorithm
Multiparty Algorithm is implemented to restrict the user who is generating the association rules for not accessing the private attributes of the customer. Options are provided for displaying the transaction details present in the databases without showing the private attributes of the customer.

IV. EXPERIMENT RESULTS

The experiment is done with the below design considerations.

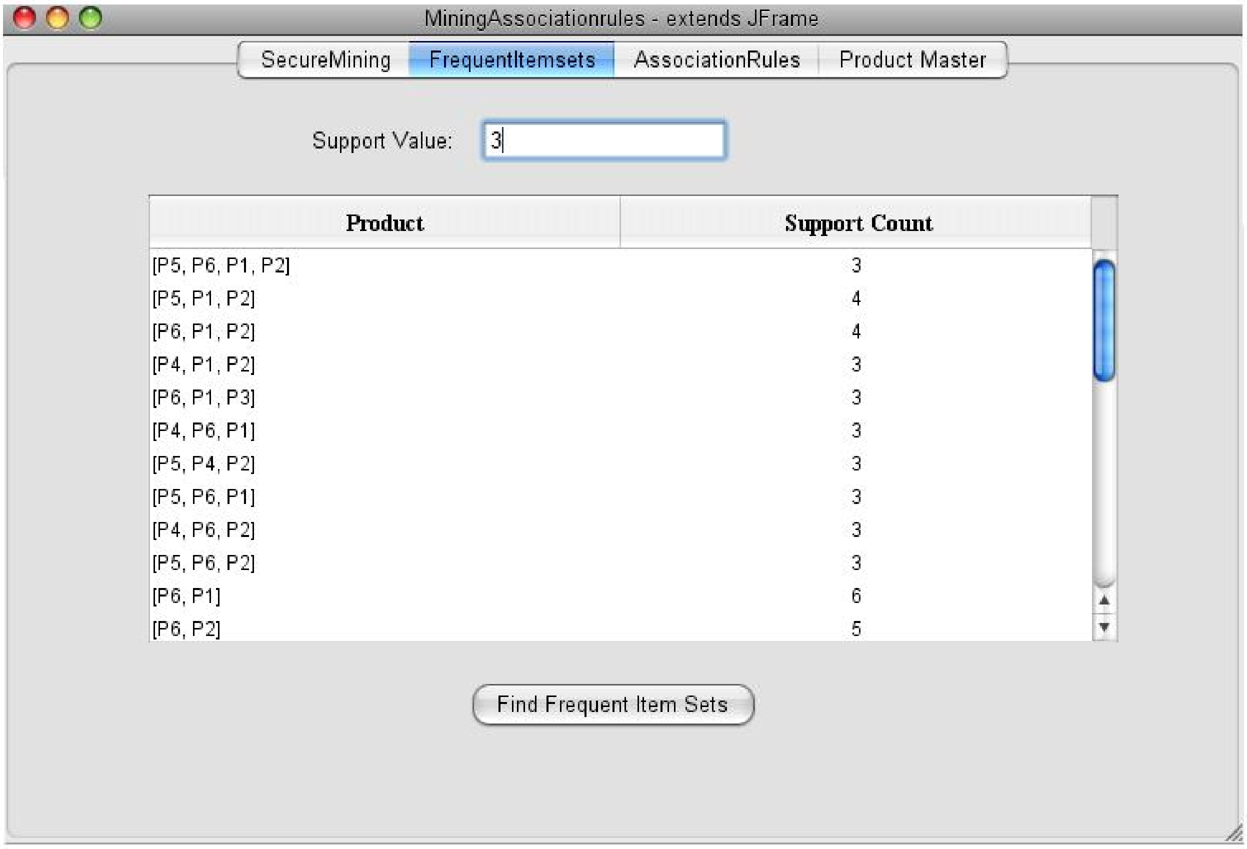
a) System is designed with static no of databases those are vertically distributed. Some databases are having some attributes of the relation R with key (transaction Id, Date) and the other databases are storing some of the attributes of the relation R along with key (transaction Id, Date) and private attributes of the customer. The transaction date and transaction id are used as key values to join the tables which are spread across databases.
b) Static no of product attributes are considered for system designing.
c) Test data are manually loaded into the databases as input.

Apriori Algorithm is implemented for finding the Frequent Item Sets from vertically distributed databases and the Association Rules are generated from Frequent Item Sets found and grouped the association rules whose confidence is greater than minimum threshold confidence. Since Databases are vertically distributed, the data's are grouped and combined before finding Frequent Item Sets and Association Rules.

The module Frequent Item Sets is tested by keying in the minimum support count. The item sets that are frequent are computed and whose support count is greater than or equal to entered minimum support count and displayed. The module Association Rules is tested by keying in the minimum confidence. The strong Association Rules that are found are grouped whose confidence is greater than or equal to entered minimum confidence threshold. To present the transaction details, options are listed in the system. The transaction details presentation excludes the private attributes of the customer.

All experiments were implemented in java JDK1.6 and were executed on a personal computer with the following configuration details such as Intel(R) Core(TM) i3-3217U personal computer with a 1.80 GHz CPU, 4 GB RAM and 64-bitoperating system Windows 7 Professional SP1.

We have tested the Frequent Item Sets module by specifying the minimum support count. We have examined that, no of Frequent Item Sets and execution time increases with lower support threshold and decreases with higher support threshold.Fig. 2. shows the Frequent Item Sets along with its support count whose values are greater than or equal to minimum support count 3which is captured from the users.

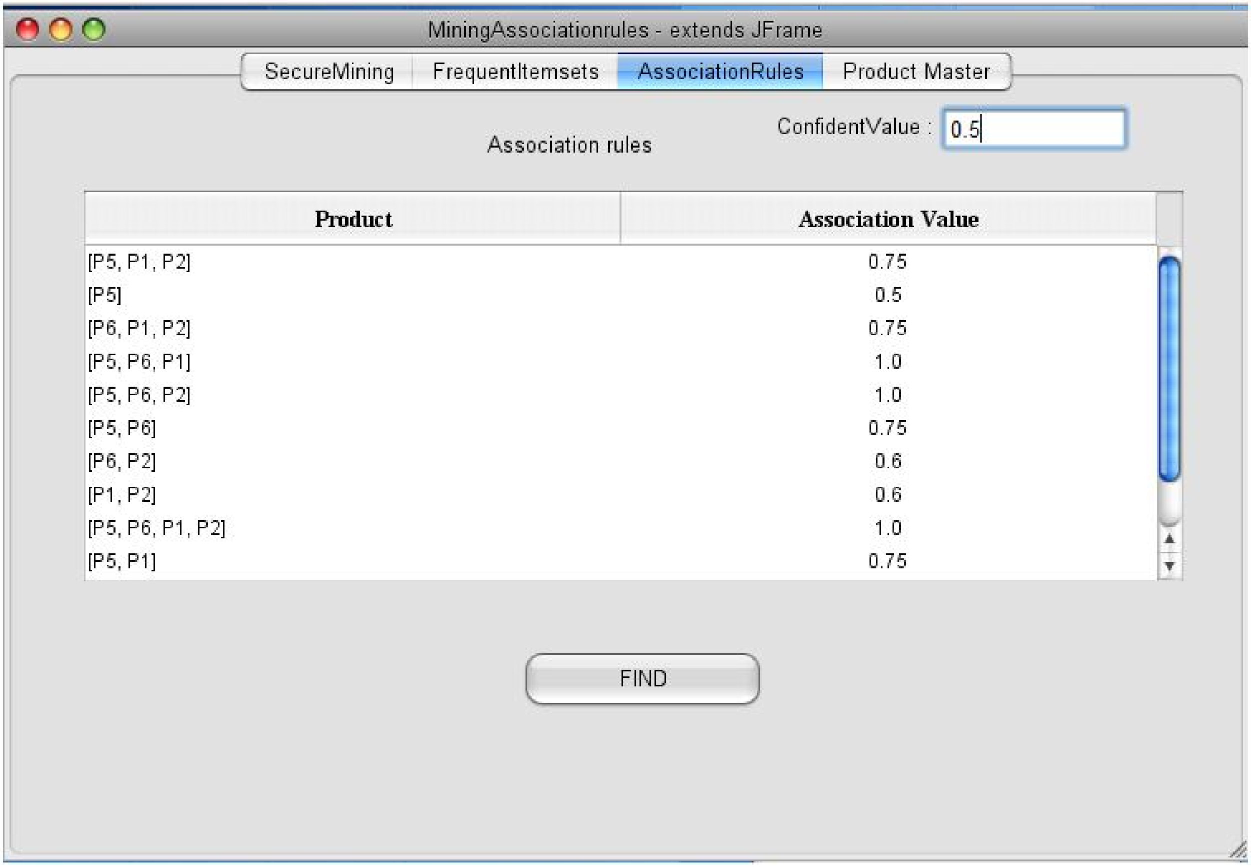


Support Value: 3

Product	Support Count
[P5, P6, P1, P2]	3
[P5, P1, P2]	4
[P6, P1, P2]	4
[P4, P1, P2]	3
[P6, P1, P3]	3
[P4, P6, P1]	3
[P5, P4, P2]	3
[P5, P6, P1]	3
[P4, P6, P2]	3
[P5, P6, P2]	3
[P6, P1]	6
[P6, P2]	5

Fig.2. Frequent Item sets

We have tested the Association Rules module by specifying the minimum confidence. We have examined that, no of Association Rules and execution time increases with lower confidence threshold and decreases with higher confidence threshold.Fig. 3. shows the Strong Association Rules along with its confidence whose confidence is greater than or equal to minimum confidernce entered 0.5%



ConfidentValue : 0.5

Product	Association Value
[P5, P1, P2]	0.75
[P5]	0.5
[P6, P1, P2]	0.75
[P5, P6, P1]	1.0
[P5, P6, P2]	1.0
[P5, P6]	0.75
[P6, P2]	0.6
[P1, P2]	0.6
[P5, P6, P1, P2]	1.0
[P5, P1]	0.75

Fig. 3. Association Rules

We have checked the total computation time for finding the Association Rules using the Apriori Algorithm. The computation time was measured based on the parameter “No of transactions”. The measure of productivity is the execution time of the algorithms on the vertically distributed environment. Table 1 shows total execution time taken for finding association rules using Apriori Algorithm against number of transactions.

Table 1
Execution Time for different No of Transactions

Number of Transactions	Execution Time(Seconds)
1000	9
2000	22
3000	43
4000	61
5000	84

Fig.4. shows the execution time taken for finding Association Rules against the number of transactions.

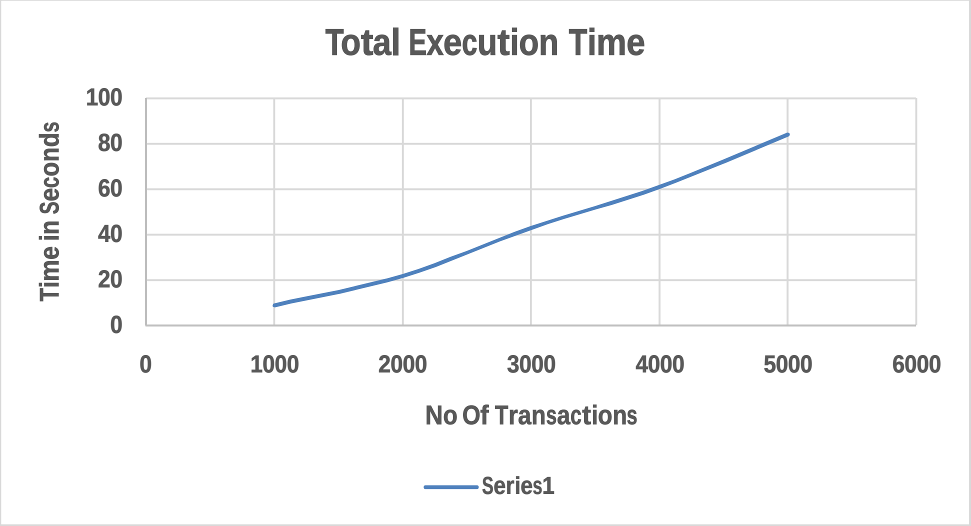


Fig. 4. Total Execution Time

V.CONCLUSION

We have proposed a protocol for secure mining of association rules in vertically distributed databases. Experiment results have been demonstrated the results of the proposed algorithms over vertically distributed databases. The algorithm performance is analyzed based on the execution time and different no of transactions. Fig. 4. depicts that the efficiency of the algorithms over vertically distributed environment. In this experiment, number of databases,number of attributes are defined as static.

VI.FUTURE WORK

For further enhancement, this study suggests the implementation of the proposed algorithm or any other advanced algorithm for handling any no of product attributes in a vertically distributed database environment.

REFERENCES

[1].Tamir Tassa, “Secure Mining of Association Rules in Horizontally Distributed Databases” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014

[2].R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc 20th Int’l Conference. Very Large Data Bases (VLDB), pp. 487-499, 1994.

[3].A. Ben-David, N. Nisan and B. Pinkas, “FairplayMP - A System for Secure Multi-Party Computation,” Proc. 15th ACM Conference. Computer and Comm. Security (CCS), pp. 257-266, 2008.

[4].M. Kantarcioglu and C. Clifton, “Privacy - Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data ,”IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037,September 2004

[5].J. Vaidya and C. Clifton, “Privacy Preserving Association Rule Mining in Vertically Partitioned Data,” Proc. Eighth ACM SIGKDD Int’l Conference Knowledge Discovery and Data Mining (KDD), pp. 639-644, 2002.

[6].Vertical Partitioning Impact on Performance and Manageability of Distributed Database systems (A Comparative study of some vertical partitioning algorithms) (2006) by Hassan I. Abdalla, F. Marir 18th National computer conference 2006.

[7].R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993

[8].M. Houtsma and A. Swami. Set-Oriented Mining of Association Rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.

[9].D.Kerana Hanirex, Dr.K.P.Kaliyamurthie,” Mining Frequent Item sets Using Genetic Algorithm”, Middle-East Journal of Scientific Research, 19 (6): 807-810, 2014.

[10].Kerana Hanirex.D, Dr.K.P.Kaliyamurthie, ”Finding the Dominating Amino Acids in Dengue Virus Type1 Study on mining frequent item sets”, 4(3): (B);880 – 89, Int. Journal of Pharma and Bio Sciences, July, 2013.

[11].D.Kerana Hanirex., Dr.K.P.Kaliyamurthie, “Multi-classification Approach for Detecting Thyroid Attacks”, IJPBS, 4(3), (B) 1246 – 1251, July (2013).

[12].Kerana Hanirex.D, “An Efficient TDTR Algorithm for Mining Frequent Itemsets”, International Journal of Electronics and Computer Science Engineering, 2(1):251- 256;2012.

[13].D.Kerana Hanirex,”A Comparative Analysis on Mining Frequent Itemsets”, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), Volume1, Issue 10, P68-71, 2012.

[14].D.Kerana Hanirex, Dr.A.Kumaravel, ”An Efficient Partition and Two Dimensional Approach for Mining Frequent Itemsets”, International Journal of Technological Synthesis and Analysis (IJTSA), Volume 1, Issue 1 ,P14-17, 2012.

[15].Kerana Hanirex.D, Dr.M.A.Dorai Rengaswamy,”Efficient Algorithm for Mining Frequent Itemsets using Clustering techniques”,,IJCSE, 3(3):1028- 1032;2011.

[16].D.Kerana Hanirex ,”Association Rule Mining in Distributed Database System”, International Journal of Computer Science and Mobile Computing(IJCSMC), Vol3,Iss 4,pg 727-732,2014.

[17].P.Kalaivani, D.Kerana Hanirex, Dr.K.P.Kaliyamurthie, “Survey on Secure Mining of Association Rules in Distributed Databases”, IJRITCC, Vol 3 Iss 3:2321-8169, 2015.