International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 178– 183

A Hybrid of Apriori and FP Tree on Association Rule Mining

Vidisha H. Zodape (MTech. III[rd]sem, student)
CSE Department, PIET
Nagpur, Maharashtra, India
Email: vidisha.zodape@gmail.com

Prof. Leena H. Patil (Lecturer)
CSE Department, PIET
Nagpur, Maharashtra, India
Email: harshleena23@rediffmail.com

Abstract-Data mining is the extraction of interested information where sometimes information may be distributed at various databases. In such conditions collection of information which is of most interest of user is a challenge. This research work focuses on generation of association rule on vertically distributed database. Here a hybrid of Apriori algorithm and Frequent Pattern Tress algorithm is used to form a strong association rules. The databases which hold same structure but different objects are analogous databases. There are many data mining algorithms out of which Apriori and FP tree are the topmost algorithms. By using hybrid we can generated a strong association rule.

Keywords: *Data Mining, Distributed Database,, FP Tree, Apriori*

I. INTRODUCTION

The Web is a vast, volatile and mostly amorphous data repository, which stores incredible amount of data, and also enhance the complexity of how to deal with the information stored in database. Users wish for the tool/search engine which will provide relevant information. Service providers will have to find the techniques to create the web site by minimizing the load to best serve the siteto the different users. Business analyst wants the tool to analyze the behaviour of consumer needs.Mining is the process of finding out what users are looking for on the internet, some are interested in document file, and some users are interested in media file or images. This is the technique to find out the interesting usage pattern and best serve the information to the user. Here the method is introduced to form association rule [7] using combined apriori and FP Tree.

There are many approaches developed for secure data mining, data distribution, data modification, mining algorithm, information or rule hiding and privacy preserving. In distributed database some research is done on horizontally distributed database where different database records is stored in different place and some on Vertically distributed database where all values of different attributes stored in different place. Modification used in to modify the real values of a database that needs to be revealing to the public and in this way to ensure high security. It is important that a data modification techniqueshould be in production with the privacy policy used by anorganization. Data modification is done for data mining algorithms. Various data mining algorithm are designed. In information hiding or rule hiding the uninterested information or generated rule hiding is done. In privacy preservation selective modification of information is done to achieve higher modified information so that it should not be revealed.

Data mining techniques have been introduced successfully to retrieve knowledge in order to support a variety of domains marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine the data by protecting the private database of user. Most organizations want information about individuals for their own specific needs. However, different units within an organization themselves share the information. In such cases, in each they must be sure that the privacy of the individual is not violated or that sensitive business information is not revealed. In order to provide security, records can be modified before the records are shared with anyone who is not permitted directly to access the data. This can be done by deleting from the dataset some identity fields, such as name and passport number in passenger information record.

II. RELATED WORK

[1]Focuses on web usage mining. As web is mostly amorphous data repository, and also enhance the complexity of dealing with the information from the different opinion of view, users, web service providers and business analyst. They have used apriori and improved FP tree to find association rule. Apriori -the classical mining algorithm is a way to find out certain potential, regular knowledge from the massive ones. Apriori algorithm [13] is the mining of frequent item set and association rule learning [11] over transactional databases. It scans the frequent item sets by scanning the database until those items appear often in database. This is used to find the association rule[7].The FP-Tree Algorithm, is an another way to find frequent patterns without utilising candidate generations[15], therefore improving performance. It uses a divide-and-conquer strategy. The central part of this method is the usage

frequent-pattern tree (FP-tree), which keeps the piece set association information.
In simple words, this algorithm works as follows:

- It compresses the input database and gets an FPtreeinstance to represent common items.
- It then divides the compressed database into a set ofconditional databases, each one associated with onecommon pattern.
- Eventually, each database is extract separately.

Using this scheme, the FP-Tree decreases the cost to scan database looking for small patterns recursively and then combine them in the long common patterns, in large databases.

[2]Considered the applications on business environment, its benefits are defined by collaboration, team efforts and partnership, rather than individual efforts. So the collaboration is most important because it brings mutual benefit.Sometimes, collaboration even occurs among competitors, or among companies that will givethem an advantage over other competitors.

For this kind of collaboration, where all users want to share the data but want to secure the private data the method used namely Session based Secured Multiparty Collaborative Data Computation (SSMCDM). In this all participants are the authenticated users participating in data mining process and there is trusted third party which will provide session for the participants.

Figure 2.1 shows the framework of SSMCDM. Multi-session elements are used here that maintains more number of sessions for one or more participants in parallel performing various data mining tasks.

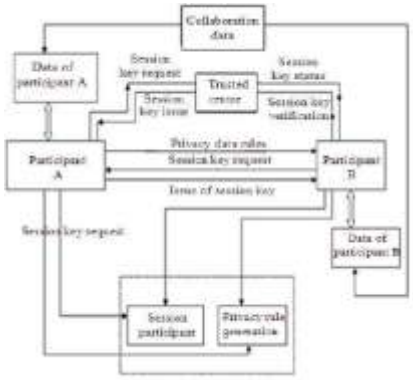


Figure 2.1: Framework of secured multiparty computation process of collaborative data

The occurrences and multi-session elements work together to maintain the private dataof participant involved in single or multiple session ofmining task simultaneously. In addition, the occurrence generation along with rule generation checks the internal deny user, who tries to retrive data in unauthorized sessions.

[3] Findsassociation rule with minimum support at least s and minimum confidence at least c and minimizing the information disclose about the private database. They have proposed FDM [8] algorithm to securely mine the association rule on horizontally partitioned database.Initially all parties already having the locally s frequent itemset. Then the candidate set is generated using apriorialgorithm. Local pruning is done to compute support and count. And then mining result is broadcasted. Here UNIFI-KC protocol is used to provide the privacy. First all parties will add to their private dataset a fake item set to hide the original size of database. Then all parties together will compute encryption of private subset followed by the computing encryptions of each itemset of subset. Then the union of subset is computed. Finally fake itemset is removed by decryption.

[4] Worked on finding rules on vertically partitioned database where the items are distributed and single entity is split over the different sites.

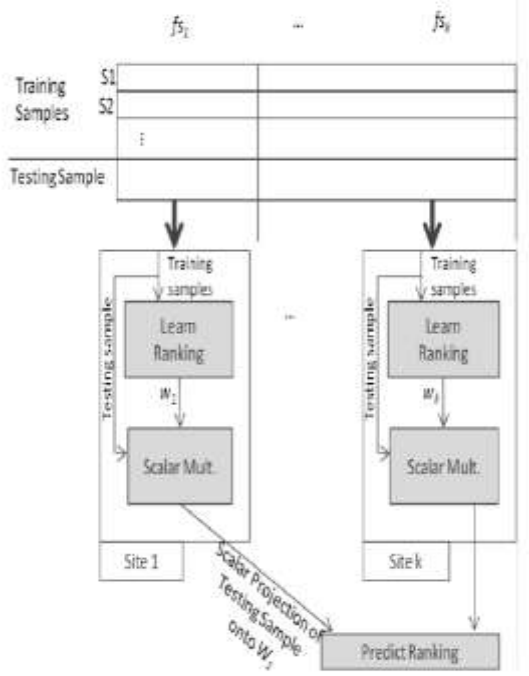


Figure 4.1: Proposed Privacy Preserving Ranking

For privacy preserving, distributed ranking is applied on individual sites and the results are merged in order to get final ranking result (Figure 4.1).The scalar projection of entities is transferred to common sites based on weight vector. Support vector machines optimization is applied to find minimum weighted vector. This is used to find the vector projection and the values are then sent to the coordinate to combine results and ranking by adding all scalar projections from all sites.

[5]Proposed new protocols which allow distributed subgroup discovery while providing security of the individual databases. Property describes a prototypical implementation and present experiments that demonstrate the feasibility of the approach. Subgroup discovery technique is used in applications based on fraud detection or clinical studies. It gives top k patterns,which uses quality of functions which is based on relative high accuracy of itemsets. Subgroup discovery [9] is very potential to provide more understanding than numerical methods like SVM or neural network. In supervised descriptive itemsets are sometimes subsumed with the classical techniques. But subgroup discovery does not provide privacy for multi party so cannot secure the information leak.

[6]Proposed commutative encryption algorithm for privacy preserving. Privacy may restrict parties to share information where data is distributed at various place. This paper gives cryptographic techniques to mine association rules on horizontally distributed database. Commutative cryptography has two phases. Figure 6.1 shows first phase. In this each party encrypts its itemset and pass it to other. This is then passed to other party to encrypt itemset and so on until all parties have encrypted all itemsets. Then common party will receive it to remove duplicates. Finally each party will decrypt each itemset and the result is common itemset.

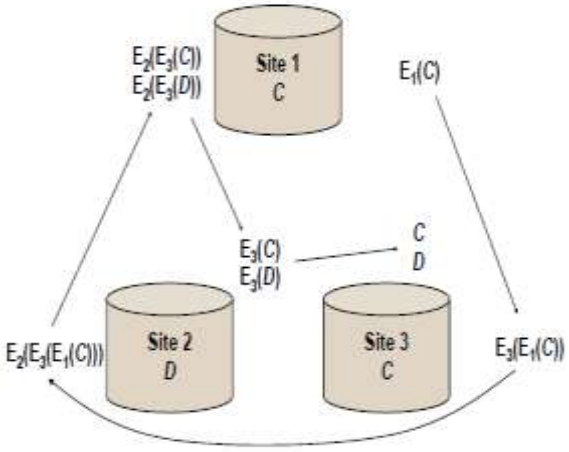


Figure 6.1: Determining Global Candidate Set

Figure 6.2 shows second phase.All locally frequent itemsets are experimented to see if they are globally s frequent. When common itemset is computed first site will choose any value X, and the exceeded amount of support is added to X. This is passed to other party and same is repeated everyone add exceeded support. The result is then compared, if it exceeds value X then it is globally supported.

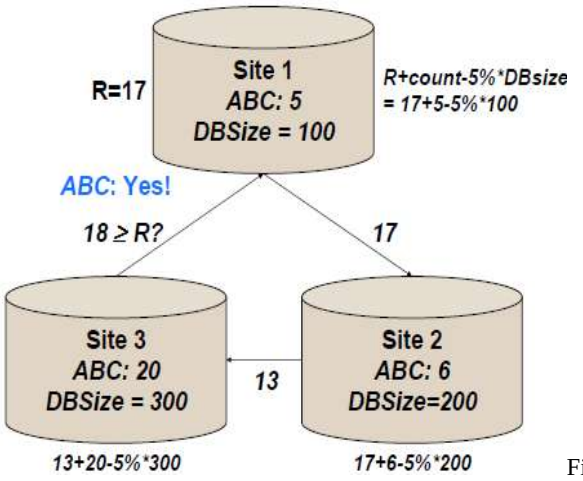


Figure 6.2: Determining If Itemset Exceeds Threshold

III. PROPOSED WORK

Here we are using a hybrid of apriori an FP tree, where firstly apriori is applied on the dataset and then the output of this will be given as the input to the Fptree. The dataset used is Crime dataset in which the information about the criminals and the crime done by them is recorded.

A. Apriori algorithm:

As we know apriori is the best algorithm to find the association rules. For example in search engine like google when we type a word we get many words which

 ISSN: 2321-8169
178– 183

are frequently associated with it that user type after that word. Let's see how apriori work with market basket example.

Original Table:

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T3	{Mango, Apples, Key-chain, Eggs}
T4	{Mango, Umbrella, Corns, Key-chain, Yo-Yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

Lets consider,
M= Mango
O= Onion
And so on..
So new table is:

Transaction ID	Items Bought
T1	{M,O,N,K,E,Y}
T2	{D,O,N,K,E,Y}
T3	{M,A,K,E}
T4	{M,U,C,K,Y}
T5	{C,O,O,K,I,E}

Step 1: Count the number of transaction for each item

Items	No. of transactions
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

Step 2: Now as per the golden rule: an item/itemset is frequently bought if it is bought at least 60% of times. So in this step we will remove items which occur less than 3 times.

Items	Number of transaction
M	3
O	3
K	5
E	4
Y	3

Step 3: Then the self-joining will be done on above table.

Item pairs
MO
MK
ME
MY
OK
OE
OY
KE
KY
EY

Step 4: Now occurrence of pairs will be counted form original table.

Item pairs	No. of transaction
MO	1
MK	3
ME	2
MY	2
OK	3
OE	3
OY	2
KE	4
KY	3
EY	2

Step 5: Remove all item pairs which occurs less than 3

Item pairs	No. of transaction
MK	3
OK	3
OE	3
KE	4
KY	3

Step 6: Again apply self-joining on above table.

Item pairs	No. of transaction
OKE	3
KEY	2

Step 7: Finally when we apply self-joining on above table we got O, K, E as final output.

B. FP tree:

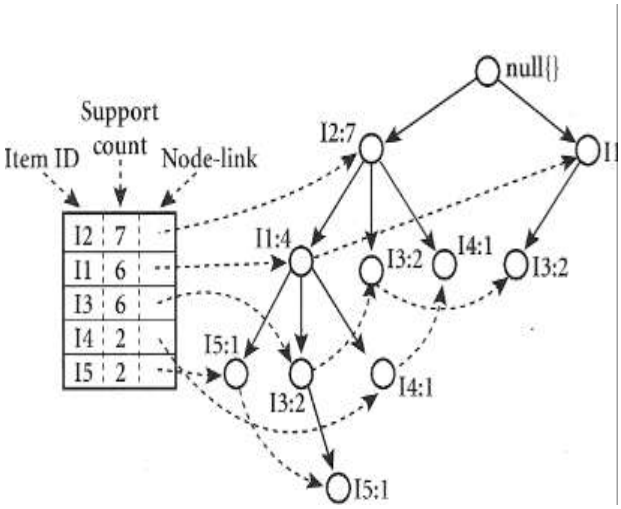
FP growth is used to construct FP tree which is the mining of frequent pattern. FP tree provides compressed dataset. It also avoids repeatedly database scanning. The working is as follows:

Firstly it scans database and finds the support for each item. Then items are removed which are not frequent. Sort other items in descending order based on counter value. Next it reads one transaction at a time and plots it on tree.

Transaction dataset

TID	Items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Now starting with reading each transaction it starts plotting tree.



Item	Conditional pattern base
I5	{(I2 I1 : 1), (I2 I1 I3 : 1)}
I4	{(I2 I1 : 1), (I2 : 1)}
I3	{(I2 I1 : 2), (I2 : 2), (I1 : 2)}
I1	{(I2 : 4)}

Item	Conditional FP-tree
I5	{I2 : 2, I1 : 2}
I4	{I2 : 2}
I3	{I2 : 4, I1 : 2} {I1 : 2}
I1	{I2 : 4}

Item	FP generated
I5	I2 I5 : 2, I1 I5 : 2, I2 I1 I5 : 2
I4	I2 I4 : 2
I3	I2 I3 : 4, I1 I3 : 2, I2 I1 I3 : 2
I1	I2 I1 : 4

C. Conclusion

The present work focuses on generation of association rules using combination of apriori algorithm and FP tree algorithm. Association rule is finding the correlation between the objects/items, association, frequent pattern in relational, transactional databases. These rules are formed by finding minimum support and minimum confidence, which helps to find most relates objects/items.As we know the data mining algorithms Apriori and FP tree both have some disadvantages. Apriori needs candidate generation which is costly mainly for large datasets. Also it needs many times to scan the database. And FP tree cannot generate good candidate set. To solve these drawbacks if we use them in combination it will form best association rule. Further we can apply security on private databases to protect sensitive information. Also we can provide security to the rules generated.

REFRENCES

[1] Ms.Monalsaxena “Association rulesMining Using Improved FrequentPattern Tree Algorithm”, International Journal of Computing,Communications and Networking, Volume 2, No.4, October - December 2013

[2] M.J. Freedman, K. Nissim, and B.Pinkas,“Privacy PreservedCollaborativeSecure Multiparty Data Mining,” Proc. Int’l Conf. Theory and Applicationsof Cryptographic Techniques (EUROCRYPT), pp. 1-19, 2012.

[3] T.Tassa and E. Gudes, “Secure Mining of Association Rules in HorizontallyDistributed Databases,” IEEETrans. Database Systems, vol. 37, 2014.

[4] J. Vaidya and C. Clifton, “Privacy Preserving Association Rule Mining in Vertically Partitioned Data,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 639- 644, 2012.

[5] H. Grosskreutz, B. Lemmen, and S. R€uping, “Secure Distributed Subgroup Discovery in Horizontally Partitioned Data,” Trans. Data Privacy, vol. 4, no. 3, pp. 147-165, 2011.

[6] T. Tassa and D. Cohen, “Privacy preseving distributed mining ofassociationrules on horizontally partitioned data ,” IEEE Trans.Knowledge and DataEng., vol. 25, no. 2, pp. 311-324, Feb. 2013.

[7] A R “Fast Algorithms for Mining Association Rules", Sep 12-15 1994, Chile, 487-99, pdf, 1-55860-153-9.
[8] MannilaH,"Efficient algorithms for discovering association rules mining." conference Knowledge Discovery in Databases (SIGKDD). 181-83.
[9] Tan, P. N., M. St., V. Kumar, “Introduction to web Mining”, Addison-Wesley, 2013, 769pp.
[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, 2nd ed. San Mateo.
[11] Huang, H., Wu, X..Association analysis with one scans of web data bases. Paper submitted at the IEEE On Data Mining, Japan.
[12] R. Jin “An Efficient Implementation of Apriori Association web mining,” Proc. Workshop on High Performance Data webMining, Apr. 2011.
[13] J. H and M. Kaber, “association mining:” 2014.
[14] Han J “Mining frequent patterns without candidaterules mining technique," in the national seminar of theinternational web of data, ACM Press, pp. 4-11-2004
[15] E-H. Han, G. Caryopsis “Scalable Data web miningfor Association web Rules,” IEEE Trans. Eng., vol. 12,no. 3, July 2012.
[16] Brin S., R. Mot, J.D. Ullman, “web item set countingand implication rules
[17] Association mining in data base”, in Proceedings ofthe ACM SIGMOD International Conference onManagement of Data, pp.289-294, 1999.
[18] Masseglia F., “Using Data Mining Techniques onWeb Access Logs to Dynamically Improve HypertextStructure language”, In ACCM Web Letters, Vol. 10 No.9, pp.13-19,