An Efficient Rare Interesting Item Set Mining using Modified MCCFP-Growth

Patel Rina N.[1], Prof. Khushboo Trivedi[2]
[1]Dept of Computer Science and Engineering, Asst. Prof.,
[2]Dept. of Information Technology
Parul Institute of Engineering and Technology, Vadodara, India.

ABSTRACT

Rare association rule is an association rule consisting of rare items. It is difficult to mine rare association rules with a single minimum support constraint because low minsup can result in generating too many rules in which some of them are uninteresting. In this paper ‘maximum constraint model’ which uses multiple minsup constraint has been proposed and extended to Apriori approach for mining frequent patterns. This model is efficient ,the Apriori-like approach raises performance problems. FP-growth like approach utilizes the prior knowledge provided by the user at the time of input and also discovers frequent patterns with good performance. The MCCFP-Growth approach which takes so much time during inserting Interesting and Non-Interesting Items. Hence, we propose Modified MCCFP-Growth will take less time because of only insering Rare Interesting item with the Support Value is less than to its MSI Value.The Experimental result shows the processing time for overall process will be decreased and system will become efficient without affecting the number of rules generated.In future,there is improvement in Modified MCCFP-Growth in terms of Memory.

Keywords— *Data mining, FP-tree, Maximum Constraint Model, Minimum item support, Rare association rules.*

I. INTRODUCTION

Data mining or knowledge discovery in databases represents techniques for discovering knowledge patterns hidden in large databases. Many data mining approaches are being used to extract interesting knowledge like association rule mining. Data mining is seen as an increasingly important tool by modern business to transform data into the business intelligence that giving the informational advantages(Agrawa. R. and Srikant. R. , 1994). Data mining is currently used in the wide range of profiling practices, such as scientific discovery, marketing, fraud detection and surveillance(Rachna Somkumar,2012).Association rule mining is most important data technique which discovers interesting associations among the items in a dataset. An association rule is an expression of the form X→Y,where X, Y are itemsets. It shows the relationship between the items X and Y. and The fraction of transactions containing X also containing Y, i.e., $P(Y|X)=P(X\cup Y)/P(X)$ is called the confidence of the rule. and support (sup) of the rule is the fraction of the transactions that contain all items both in X and Y, i.e., $sup(X\rightarrow Y) = P(X\cup Y)$ (R.Uday Kiran and P.krishna Reddy,2010).

A rare association rules refers to an association rule forming between either frequent and rare items that used for the knowledge in rare associations (Rakesh,Agrawal,Tomasz,lmielinski,ArunSwami,1993).

Association rule mining technique is most used application of data mining in retail stores and even for business intelligence. Generation of association rules can be done after find the frequent patterns from the available dataset. Frequent patterns have the property of high support. User defines the minimum support criteria for the item to be frequent. If the itemset satisfy the minimum support criteria than and only than it can be there in frequent patterns. There exist some items with low support but that item have high confidence. Mining of this type of items is call rare itemset mining.

In Minimum Constraint Model, each item is specified with a support constraint,called minimum item support(MIS).and in Maximum Constraint model pattern be a frequent and it must satisfy only the lowest MIS value among all items.

II. RELATED WORKS

Apriori algorithm

Apriori algorithm has been proposed in (Agrawal et al., 1993; Agrawal and Srikanth, 1994) for finding frequent itemsets. Apriori is used for learning association rules.It is designed to operate on databases containing transaction. It is more efficient during the candidate generation process(Agrawa. R. and Srikant. R. , 1994). It uses the pruning techniques to avoid the measuring certain itemsets, while guaranteeing completeness. There are two processes to find out all large itemsets from the database in apriori algorithm.one is the candidate itemsets are generated,after that the database is scanned to check actual support count of the each item is calculated and the large 1-itemsets are generated by pruning those itemsets whose support are below the pre-define threshold. In each pass only the those candidate itemsets that include the same specified number of the items are generated and checked. The candidate k itemsets are generated after the (k-1)th passes over the database by joining the frequent k-1 itemsets. All the candidate k-itemsets are purned by check their sub (k-1)-itemsets, this k-itemsets candidate is pruned out because it has no hope to be frequent according the apriori property.This algorithm is based on iterative level wise serch for frequent pattern generation. It uses a single minsup value for all levels to extract the frequent itemsets, so the algorithm generates all candidates itemsets in that level.A Candidate k-itemset is an itemset having 'k' number of items. A candidate k-itemset is said to be frequent if the support of the subset of candidate k-itemsets is greater than or equal to the user-specified minsup threshold. This algorithm is suitable for find out the frequent itemsets and not the rare itemsets(R.Uday Kiran and P.krishna Reddy ,2010)(Kanimozhi Selvi Chenniangirivalsu Sadhasivam and Tamilarasi Angamuthu ,2011).

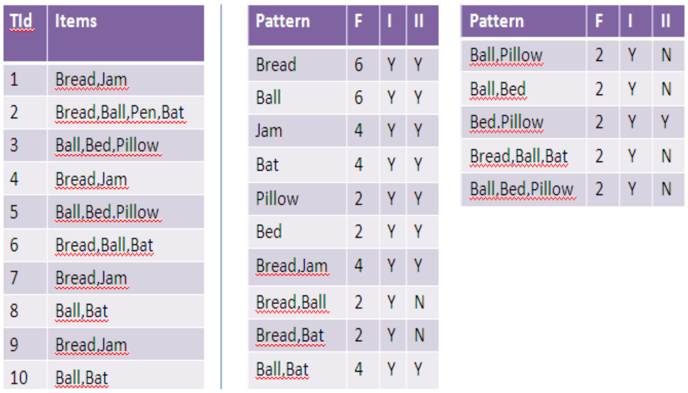
This algorithm inherits the drawback of scanning the whole databases many time. It also take the much time, space and memory to the candidate generation process. Based on the this algorithm, many new algorithms were designed with some modification or improvements. There were two approaches: (1)Reduce the number of the passes over the whole database or replacing the whole database with only part of it based on the current frequent itemsets. (2)To explore different kinds of pruning techniques to make the number of candidate itemsets much smaller.

MSApriori algorithm

In the MSApriori algorithm,it is an extension of Apriori algorithm so it is called MSApriori algorithm and it has been proposed in (Liu et al., 1999) which attempts to discover frequent itemsets involving rare items.This algorithm assigns a minsup value known as MIS for each item and frequent itemsets are generated if an itemset satisfies the lowest MIS value among the respective items. Using this method derives the MIS values for items based on their support percentage. Here frequent items are assigned with a higher MIS value whereas rare items are assigned with a lower MIS value. So, MSApriori algorithm having rare itemset problem(R.Uday Kiran and P.krishna Reddy ,2010).

Maximum Constraint Based Conditional Frequent Pattern-Growth(MCCFP-Growth)

Here the Fig.1 shows all transactional dataset Tran and items with MIS values as an input parameters. Using the items MIS values as prior knowledge and discovers frequent patterns with a single scan on the transactional dataset. The steps for this approach is as follows:(1) Constructing of a tree, called MIS-tree. (2)Generating compact MIS-tree from MIS-tree.(3)Mining compact MIS-tree using conditional pattern bases to discover complete set of frequent patterns.



TId	Items
1	Bread,Jam
2	Bread,Ball,Pen,Bat
3	Ball,Bed,Pillow
4	Bread,Jam
5	Ball,Bed,Pillow
6	Bread,Ball,Bat
7	Bread,Jam
8	Ball,Bat
9	Bread,Jam
10	Ball,Bat

Pattern	F	I	II
Bread	6	Y	Y
Ball	6	Y	Y
Jam	4	Y	Y
Bat	4	Y	Y
Pillow	2	Y	Y
Bed	2	Y	Y
Bread,Jam	4	Y	Y
Bread,Ball	2	Y	N
Bread,Bat	2	Y	N
Ball,Bat	4	Y	Y

Pattern	F	I	II
Ball,Pillow	2	Y	N
Ball,Bed	2	Y	N
Bed,Pillow	2	Y	Y
Bread,Ball,Bat	2	Y	N
Ball,Bed,Pillow	2	Y	N

Fig.1: (a) Transactional dataset and (b) Patterns having support count(or support) greater than or equal to 2.F represents the support count of the pattern.The term 'Y' and 'N' in these columns corresponds to frequent patterns generated and have not generated in the respective approaches.(R.Uday Kiran and Polepalli Krishna Reddy, 2012)

Structure of MIS-tree: It consist of two components: (1)MIS-list and (2) Prefix-tree. The MIS-list is a list having three fields item name(item), frequency(S), minimum item support (MIS).so,the structure of the prefix-tree is same as FP-tree. Using the transactional dataset which is shown in fig.1(a), given MIS values for the items Bread,Ball,Pen,Jam,Bat,pillow and Bed be 4,4,3,3,3,2 and 2 respectively(Ya-han Hu and Yen-Liang Chen,2004),(R.Uday Kiran and P.Krishna Reddy,2009).

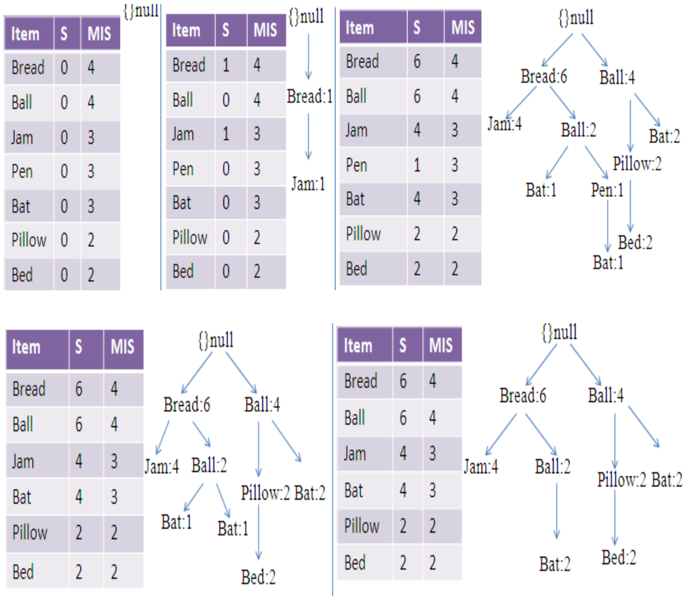
1. Construction of MIS-tree

First of all you have to arrange all items in descending order of with respect to their MIS values. after that sorted list of items L.={ Bread, Ball, Pen, Jam, Bat, Pillow, Bed}.In the L order, insert each item into the MIS-list with f=0 and MIS is equivalent to their respective MIS values.

For creating the prefix-tree, first we have to create a root node and label it as “null”. The MIS-tree created before scanning the transactional dataset is shown in fig.2(a).The first transaction”1:”Bread,Jam” containing two items in the transactional dataset is scanned in L order,i.e.,{Bread, Jam},and their frequencies are updated by 1 in the MIS-list.for this transaction,in L order, a branch is created in the prefix-tree as in FP-growth.The updated MIS-tree after scanning first transaction is shown in figure2(b).Every transaction is scanned and MIS-tree is updated. The updated MIS-tree after scanning every transaction is shown in Figure2(c). For tree traversal,node links are maintained as in FP-tree.

Deriving Compact MIS-tree

During the downward closure property,items which have support less than their respective MIS values cannot generate any frequent pattern.so,such items can be pruned from the MIS-tree so that it is compact.



Item	S	MIS
Bread	0	4
Ball	0	4
Jam	0	3
Pen	0	3
Bat	0	3
Pillow	0	2
Bed	0	2

Item	S	MIS
Bread	1	4
Ball	0	4
Jam	1	3
Pen	0	3
Bat	0	3
Pillow	0	2
Bed	0	2

Item	S	MIS
Bread	6	4
Ball	6	4
Jam	4	3
Pen	1	3
Bat	4	3
Pillow	2	2
Bed	2	2

Item	S	MIS
Bread	6	4
Ball	6	4
Jam	4	3
Bat	4	3
Pillow	2	2
Bed	2	2

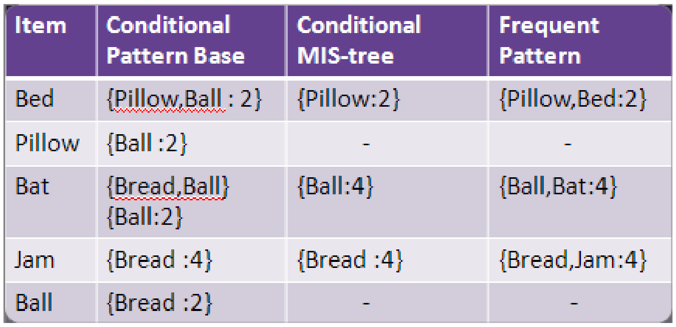
Item	S	MIS
Bread	6	4
Ball	6	4
Jam	4	3
Bat	4	3
Pillow	2	2
Bed	2	2

Fig. 2 : Construction of compact MIS-tree. (a)MIS-tree before scanning the transactional dataset (b)MIS-tree after scanning first transaction (c)MIS-tree after scanning every transaction (d)MIS-tree after pruning item’Pen’(e)compact MIS-tree derived after tree-merging operation.(R.Uday Kiran and Polepalli Krishna Reddy,2012)

Mining Frequent Patterns from compact MIS-tree

In the step of Mining of frequent paterns from the compact MIS-tree is same as the mining of frequent patterns from the FP-tree. The difference is that MIS value of the prefix-item or pattern from conditional pattern base of a suffix pattern is used as minsup. so,mining the frequent patterns in compact MIS-tree is similar as fp-tree.

Compact MIS-tree is shown in Fig.2(e) for mining the patterns is shown in Fig.3.suppose we take path <Ball,Pillow,Bed:2>.we choosing Bed as a suffix,and its corresponding path is <Ball,Pillow:2>.Ball is not concluded because its support is less than to their respective MIS value.



Item	Conditional Pattern Base	Conditional MIS-tree	Frequent Pattern
Bed	{Pillow,Ball : 2}	{Pillow:2}	{Pillow,Bed:2}
Pillow	{Ball :2}	-	-

Bat	{Bread,Ball}
	{Ball:2}
		{Ball:4}	{Ball,Bat:4}

Jam	{Bread :4}	{Bread :4}	{Bread,Jam:4}
Ball	{Bread :2}	-	-

Fig. 3 : Mining Frequent patterns using conditional pattern bases(R.Uday Kiran and Polepalli Krishna Reddy,2012).

III. LIMITATIONS OF MCCFP-GROWTH

Current system is using the FP-tree based approach to solve the problem. When they first create the tree at that time whole data is loaded in to the memory. And after that we need to remove the non-interesting items by pruning the tree. Thus even if the dataset is sparse algorithm will take more time because of pruning will take more time. If dataset is dense then initially we require more memory to construct the tree which contains whole dataset. Thus there is scope of improvement in the algorithms in terms of time and memory both.

IV. MODFIED MCCFP-GROWTH

The ‘rare itemset problem’ is solved using two different constrained models. First is ‘minimum constrained model’ and second is ‘maximum constrained model’. Maximum constrained model satisfy the downward closure property and that is the reason why it is of user’s interest. We focused on to use the ‘maximum constrained problem’. There exist so many approaches to address this problem. But our interest is in approach used in (Sidney Tsang,Yun Sing Koh and Gillian Dobbie

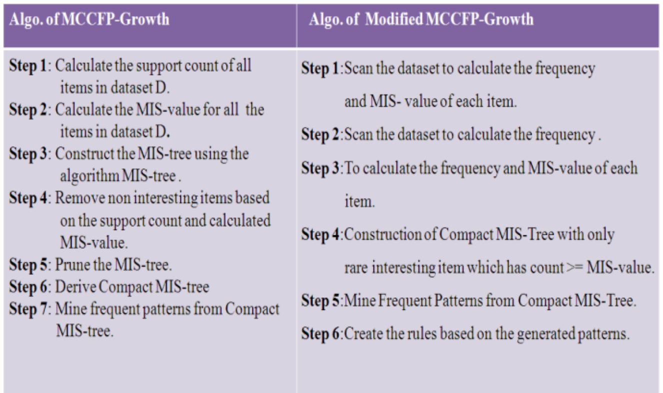
, 2013) because of its compact data structure and less time consumption while running.

Using the same approach as used in base paper. But there is a scope of improvement in the MCCFP-Growth algorithm. According to the algorithm proposed in base paper first it calculates the frequency of each and every item to calculate its MIS-value based on the equation (1). After calculating the MIS-value for all items it start building the tree using the MIS-tree algorithm. But as we have the support count of all the items we can eliminate the items which are having support count less than the calculated MIS-value before inserting the itemset in a MIS-Tree.

So we can find the items which only can be considered as rare interesting items. Thus there is no need to insert the items which are not included in the rare interesting item. Thus by just giving only rare interesting items in the input while creating the MIS-Tree we need not to perform the step of :(a) Removing items with counts less than MIS-value and Need not to prune the tree.

The advantages of proposed approach is it Takes less time because of tree pruning step is not needed and it Takes less amount of memory when constructing the MIS-tree.

Flow of Existing System and proposed System

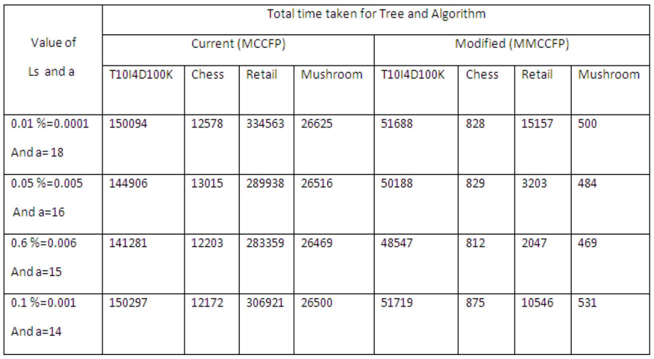


Algo. of MCCFP-Growth	Algo. of Modified MCCFP-Growth
Step 1: Calculate the support count of all items in dataset D. **Step 2**: Calculate the MIS-value for all the items in dataset D. **Step 3**: Construct the MIS-tree using the algorithm MIS-tree . **Step 4**: Remove non interesting items based on the support count and calculated MIS-value. **Step 5**: Prune the MIS-tree. **Step 6**: Derive Compact MIS-tree **Step 7**: Mine frequent patterns from Compact MIS-tree.	**Step 1**:Scan the dataset to calculate the frequency and MIS- value of each item. **Step 2**:Scan the dataset to calculate the frequency . **Step 3**:To calculate the frequency and MIS-value of each item. **Step 4**:Construction of Compact MIS-Tree with only rare interesting item which has count >= MIS-value. **Step 5**:Mine Frequent Patterns from Compact MIS-Tree. **Step 6**:Create the rules based on the generated patterns.

V. RESULTS AND EVALUATION

A. Advantages Using modified MCCFP-Growth during the experimental result even for the dense and sparse dataset the algorithm is taking less time The MCCFP-Growth (Maximum Constraint Based Conditional Frequent Pattern-Growth) also consider the Interesting and Non-Interesting items while constructing the tree.

B. The modified MCCFP-Growth algorithm will consider only the Rare Interesting items while constructing the tree from the dataset. During constructing tree, modified MCCFP-Growth algorithm discards or Prune those items which have support value is less than to its MIS value, i.e. non-interesting item. Thus the tree, which takes so much time, will be avoided. As a result the processing time for overall process will be decreased and the system will become more efficient without affecting the number of rules generated.



Value of Ls and a	Total time taken for Tree and Algorithm							
	Current (MCCFP)				Modified (MMCCFP)			
	T10I4D100K	Chess	Retail	Mushroom	T10I4D100K	Chess	Retail	Mushroom
0.01 %=0.0001 And a= 18	150094	12578	334563	26625	51688	828	15157	500
0.05 %=0.005 And a=16	144906	13015	289938	26516	50188	829	3203	484
0.6 %=0.006 And a=15	141281	12203	283359	26469	48547	812	2047	469
0.1 %=0.001 And a=14	150297	12172	306921	26500	51719	875	10546	531

C. Sparse and dense dataset are used for taking execution time.

VI. CONCLUSIONS

This paper provides brief introduction about the algorithms which is used in the area of rare association mining. From the entire available algorithm only focused on the maximum constraint based algorithm. Due to the existence of only an apriori-like approach, mining rare association rules or frequent patterns using this algorithm raises performance problems.Using modified MCCFP-Growth the processing time for overall process will be decreased and the system will become more efficient without affecting the number of rules generated. The result even for the dense (chess,mushroom) and sparse(T10I4D100K,retail)like dataset the modified MCCFP-Growth is taking less time as compare to MCCFP-Growth.The MCCFP-Growth (Maximum Constraint Based Conditional Frequent Pattern-Growth) also consider the Interesting and Non-Interesting items while constructing the tree.

The modified MCCFP-Growth algorithm will consider only the Rare Interesting items while constructing the tree from the dataset. During constructing tree, modified MCCFP-Growth algorithm discards or Prune those items which have support value is less than to its MIS value, i.e. non-interesting item. Thus the tree, which takes so much time, will be avoided. As a result the processing time for overall process will be decreased and the system will become more efficient without affecting the number of rules generated.

In future, improve the algorithm that is MCCFP-Growth in terms of memory with the help of empirical analysis.

Acknowledgement

I would like to express the deepest appreciation to Ms. Khushboo Trivedi who has guided me and for their support and motivation that they have provided. She has always been willingly present whenever I needed the slightest support from her. I would not like to miss a chance to say thank for the time that she spared for me, from her extremely busy schedule.

References

1. R.Uday Kiran and P.krishna Reddy (2010)., "Mining Rare Association Rules in the Datasets with Widely Varying Item's Frequencies",*Center of Data Engineering, International Institute of Information technology-Hyderabad pp. 49-62,* Springer.
2. Ya-han Hu and Yen-Liang Chen (2004). "Mining association rules with multiple minimum supports :a new mining algorithm and a support tuning mechanism", *Department of Information Management,,Taiwan,Roc*, Elsevier.
3. R.Uday Kiran and P.Krishna Reddy(2009). "An Improved Multiple Minimum Support Based Approch to Mine Rare association Rules",IEEE.
4. Yeong-Chyi Lee,Tzung-Pei Hong, Wen-Yang Lin (2005). "Mining association rules with multiple minimum supports using maximum constraints", *Institute of Information Engineering, i-Shou University, Kaohsiung 840,Taiwan, Roc*, Elsevier.
5. Kshat Surana,R.Uday Kiran,P.krishna Reddy (2009). "Selecting Right Interestingess Measure for rare Association Rules", Center for Data Engineering, *International Institute of Information technology-Hyderabad,Computer Society of India.*
6. Chin-chen,Chang,Yu-Chiang,Li (2005). "An Efficient Algorithm for Increment Mining of Association Rules", *Department of Information Engineering and Computer Science,feng Chia university, Taichaung, Taiwan*, IEEE.
7. Rakesh,Agrawal,Tomasz,lmielinski,ArunSwami (1993). "Mining Association Rules between Sets of Items in Large databases",*IBM Almaden research center,650 Harry Road,SanJose*,ACM.
8. Kanimozhi Selvi Chenniangirivalsu Sadhasivam and Tamilarasi Angamuthu (2011), "Mining Rare Intemset with Automated Support Thresholds",*Department of Computer Applications,Kongu Engineering college, Perundurani, Erode,Jouranl Of Computer sacience*,.
9. Sidney Tsang,Yun Sing Koh and Gillian Dobbie (2013). "Finding Interesting Rare Association Rules Using Rare Pattern Tree",*The University of Auckland,Springer-Verlag Berlin Heidelberg.*
10. Sunitha Vanamala,L.Padma sree,S.Durga Bhavani (2013). "Efficient Rare Association Rule Mining Algorithm",*International Journal of Engineering Reserch and application,*
11. Jyothi Patil,Dr.V.D.Mytri (2013). "A Fast Association rule Algorithm Based on Bitmap Computing with Multiple Minimum Supports using Max Constraints",*International Journal of Comp. Sci & Electronics Engg.,Volume1,Issue 2*, IJCSEE.
12. R.Agrawal, T. Imielinski, and A. Swami (1993). "Mining association rules between sets of items in large databases" *In Proceeding ACM SIGMOD Conference, pp. 207-216*, ACM.
13. R.Agrawal, T. Imielinski, and A. Swami (1993). "Mining association rules between sets of items in large databases" *In P. Bunemann and S. Jajodia, editors, Proceedings of the 1993 ACM SIGMOD Conference on Management of Data* . 207-216, New york, ACM.
14. Agrawa. R. and Srikant. R. (1994). "Fast algorithms for mining association rules" *In Proc. 20th Int..Conf. Very Large Data Bases, 487-499.*
15. Rachna Somkumar (2012). "A Study on various Data Mining Approaches of Association Rules", *Int.J.Comput. Sci. Eng., vol2, pp.141-144.*
16. Agrawa. R. and Srikant. R.(1994), "Fast algorithms for mining association rules", *In Proc. 20th Int..Conf. Very Large Data Bases, 487-499.*
17. Rachna Somkumar (2012), "A Study on various Data Mining Approaches of Association Rules", *Int.J.Comput. Sci. Eng., vol2, pp.141-144.*
18. R.Uday Kiran and Polepalli Krishna Reddy(2012), "An Efficient Approch to mine Rare Association Rules Using Maximum Items' Supports Constraints", *International institute of Information Technology-Hyderabad,* Springer.