AN OVERVIEW OF TECHNIQUES IN MINING ASSOCIATION RULES

W.Sarada*

B.Laksmi Prasanna*

Abstract

Association rule discovery is one of the most important techniques in the field of data mining. Key component of data mining is whether Knowledge Discovery or Knowledge Prediction, data mining takes information that was once quite difficult to detect and presents it in an easily understandable format (i.e. graphical or statistical).Advanced data mining methods rely on Relational Data, or data that can be stored and modeled easily through use of relational databases. It aims at finding interesting patterns among the databases. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. They are employed today in many application areas including web usage mining, intrusion detection and bioinformatics .The use of association rule mining technique is to describe the associations among items in a database and to identify domain knowledge hidden in large volume of data efficiently. This paper provides an overview of techniques that are used to improve the performance and efficiency of Association Rule Mining (ARM) from huge databases.

***Index Terms**: Data Mining, Domain knowledge, Association rules, Databases, Frequent Patterns*

* Assistant Professor, Dept. of Comp.Sci, R.B.V.R.R.Women's College (Autonomous), Andhra Pradesh,India

1. INTRODUCTION

Data Mining (DM) is defined as the discovery of interesting information, patterns or trends from a large database or data warehouse. Data mining is a sub process of Knowledge Discovery in Databases in which the different available data sources are analyzed using various data mining algorithms. Speaking of DM we refer to “a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval and visualization” .The two high-level goals that data miners want to achieve are prediction and description:

• Prediction is used to find patterns which can be used to project the future.

• Description represents discovered patterns to the user in a human-understandable form.

The gained knowledge is either represented as a model or generalization of the mined data. The data needed to conduct the data mining is widely available in the age of the Internet and e-commerce.

The ‘mined’ information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification.

2. ASSOCIATION RULES

Association rules mining is especially efficient for its use in the Internet. The data are easily available and mining can thereby be done quickly. Discovered rules can help online-shops in personalizing their website and cross-selling their products by making recommendations. A problem of classical association rules is that not every kind of data can be used for mining. Rules can only be derived from data containing binary data, where an item either exists in a transaction or it does not exist. When dealing with a quantitative database, no association rules can be discovered. This fact led to the invention of quantitative association rules, where the quantitative attributes are split into intervals and the single elements are either members or nonmembers of those intervals.

An association rule can be expressed as the form A -> B, where A and B are sets of items, such that the presence of A in a transaction will imply the presence of B. Two measures, support and confidence, are evaluated to determine whether a rule should be kept. The support of a rule is the fraction of the transactions that contain all the items in A and B. The confidence of a rule is the conditional probability of the occurrences of items in A and B over the occurrences of items in A. The support and the confidence of an interesting rule must be larger than or equal to a user-

specified minimum support and a minimum confidence respectively. A typical and widely-used example of association rule mining is Market Basket Analysis. The problem is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum
confidence.

Find all the rules *X* → *Y* with minimum support and confidence

- support, *s*, probability that a transaction contains X ∪ Y.
- confidence, *c*, conditional probability that a transaction having X also contains *Y*
- *Let minsup = 50%, minconf = 50%*
- *Frequent Patterns:* Bread:3, Nuts:3, Butter:4, Eggs:3, {Bread, Butter}:3
- Association rules: (many more!)
  Bread → *Butter* (60%, 100%)
  Butter→ *Bread* (60%, 75%)

Association rules are especially useful for conducting a market basket analysis where transaction data can be analyzed. Regularities in data of a supermarket for example can be found in this way. An association rule could be "If a customer
buys bread and milk, he will mostly buy butter as well". This information is very useful for business because promotion actions can be designed accordingly. Association Rules and Goals
•Find all sets of items (item-sets) that have support (number of transactions) greater than the minimum support(large item-sets).
•Use the large item-sets to generate the desired rules that have confidence greater than the minimum confidence.

2.1 About Mining of Association Rules

One of the most popular data mining approaches is the Apriori algorithm [1](Agrawal et al., 1993) which finds association rules in two steps.
1)First, all item sets *x* with support of more than the fixed threshold "minsup" are found.

2)Then, all item sets are split into left and right hand side *x* and *y* (in all possible ways) and the confidence of the rules [*x* => *y*] is calculated as $\frac{s(x \cup y)}{s(x)}$. All rules with a confidence above the confidence threshold "minconf" are returned.i.e. The problem of rule mining is divided into two sub problems:

● It is necessary to identify all combinations of items that have a transaction support above a certain threshold, called min support. We will call those sets of items that show a sufficient support large or frequent itemsets, and those not meeting the threshold small itemsets. Syntactic constraints can also be taken into consideration, for example if we are only interested in rules that contain a certain item in the antecedent or the consequent.

● After identifying the itemsets that satisfy the min support, it is important to test whether it satisfies the confidence factor *c*. Only the previously defined large itemsets have to be tested at this stage. The confidence is computed by dividing the support of the whole item set by the support of the antecedent.

After having solved the first problem in finding all relevant large itemsets, the second part is rather straightforward.

3. Association Rule Mining Algorithms

3.1Apriori algorithm

In order to discover large itemsets, the Apriori algorithm was developed as the first and nowadays best known algorithm for mining association rules. However, there are two bottlenecks in the Apriori algorithm. One is the complex frequent itemset generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database.

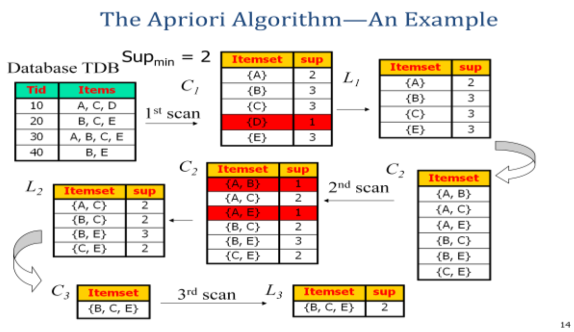

Fig-1: Item set generation

3.2Aprioritid algorithm

The AprioriTid algorithm also uses the apriori-gen function to determine the candidate itemsets before the pass begins. The interesting feature of this algorithm is that the database D is not used for counting support after the first pass. It is not necessary to use the same algorithm in all the passes over the data. Apriori still examines every transaction in the database. On the other hand, rather than scanning the database, AprioriTid scans Ck for obtaining support counts, and the size of Ck has become smaller than the size of the database.Based on these observations AprioriHybrid algorithm has been designed. This uses Apriori in the initial passes and switches to AprioriTid.

3.3 Apriorihybrid algorithm

It is not necessary to use the same algorithm in all the passes over the data.Apriori still examines every transaction in the database. On the other hand, rather than scanning the database, AprioriTid scans Ck for obtaining support counts, and the size of Ck has become smaller than the size of the database.Based on these observations AprioriHybrid algorithm has been designed Figure 7 shows the execution times for Apriori and AprioriTid for different passes. In the earlier passes, Apriori does better than AprioriTid. However, AprioriTid beats Apriori in later passes, the reason for which is as follows. Apriori and AprioriTid use the same candidate generation procedure and therefore count the same itemsets. In the later passes, the number of candidate itemsets reduces however, uses Apriori in the initial passes and switches to AprioriTid in the later passes.

The AprioriTid and AprioriHybrid have been proposed to solve the problem of apriori algorithm. From the comparison we conclude that the AprioriHybrid is better than Apriori and AprioriTid, because it reduced overall speed and improve the accuracy.

3.4 AIS algorithm

The AIS algorithm was the first algorithm proposed for mining Association rules. AIS algorithm consists of two phases. The first phase constitutes the generation of the frequent itemsets. This is followed by the generation of the confident and frequent association rules in the second phase. The drawback of the AIS algorithm is that it makes multiple passes over the database. Further more,it generate and counts too many candidate itemsets that turn out to be small, which requires more space and waste much efforts that turned out to be useless

3.5 SetM algorithm

The SETM algorithm was motivated by the desire to use SQL to compute large itemsets. Like AIS, In SETM algorithm candidate itemsets are generated on the fly as the database is scanned but counted at the end of the pass. It thus generates and counts every candidate itemset that the AIS algorithm generates. However, to use the standard SQL join operation for candidate generation, SETM separates candidate generation from counting. It saves a copy of the candidate itemset together with the TID of the generating transaction in a sequential structure. At the end of the pass, the support count of candidate itemsets is determined by sorting and aggregating this sequential structure.

3.6 FPGrowth

The FPGrowth method [7]constructs FP-tree which is a highly compact form of transaction database. It generates all frequent itemsets satisfying a given minimum support by growing a frequent pattern tree structure that stores compressed information about the frequent patterns. allows frequent itemset discovery without candidate itemset generation. In this way, FP-growth can avoid repeated database scans and also avoid the generation of a large number of candidate itemsets. This algorithm mines the complete item sets without generating candidate set and uses and uses divide and conquer technique. It works on two steps-

- Build FP-Tree
- Mining of the FP-Tree to find the frequent item sets

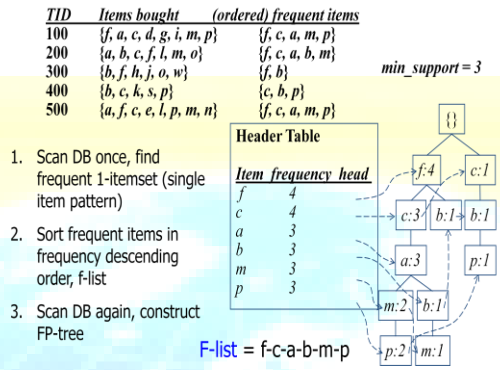

Fig-2: example of FP-tree from a Transaction Database

3.7 Reducing the number of passes

The disadvantage of Apriori algorithm made the researchers to think about new techniques to mine frequent patterns. The two main negative sides are the possible need of generating a huge number of candidates if the number of frequent 1-itemsets is high or if the size of the frequent pattern is big, the database has to be scanned twice repeatedly to match the candidates and determine the support .What if we find a way to mine the frequent patterns without candidate generation? This would be a big improvement over Apriori. That is what the frequent pattern growth (FP-growth) algorithm does [9].

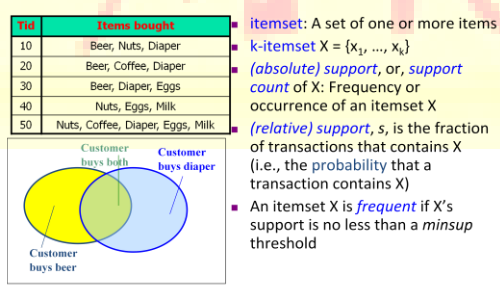

Fig-3: Frequent Patterns Example

Wang et al [10] presented PRICES, an efficient algorithm for mining association rules. Their approach reduces large itemset generation time, which dictates most of the time in generating candidates by scanning the database only once and using logical operations in the process.

Another algorithm [11] called Matrix Algorithm generates a matrix which entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets are obtained from the resulting matrix. Association rules are then mined from the frequent candidate sets.

3.8 Clustering System.

Association Rule Clustering is useful when the user desires to segment the data. Lent et al [2] proposed a Clustering Association rule in which they measure the quality of the segmentation generated by ARCS (Association Rule Clustering System) using the minimum description length (MDL) principle of encoding the clusters on several databases including noise and errors.

Scale-up experiments show that ARCS, using the BitOp algorithm scales linearly with the amount of data.

Pi et al [4] proposed a new Fuzzy Clustering Algorithm on Association Rules for Knowledge Management. A fuzzy simulation degree and simulated matrix for association rules are put forward and a new algorithm based on dynamic tree is used for implementing the fuzzy clustering. The experimental results show that this algorithm clusters the Association rules efficiently.

Gupta et al [3] recently proposed a cluster based algorithm that uses a novel approach to the insignificant transactions dynamically. During a particular pass only those clusters that seem to be statistically useful are scanned and as a consequence all insignificant tuples are filtered dynamically. The results of the algorithm show that removing false frequent items and insignificant transactions dynamically improves the performance of association rule mining.

3.8.1Partitional Clustering

Partitional Clustering is implemented by a well known method, called K Means Clustering. Here, cluster centers (centroids) are being defined and the data points are assigned to the cluster with the nearest centroid. The algorithm first computes random centers for a predefined number of clusters. Then, it recomputed the centers in order to obtain the best possible centroids.

3.8.2Hierarchical Clustering

Hierarchical clustering does not partition the data into the clusters within a single step. Rather, a series of partitions takes place. The method can be performed agglomerative or divisive. In the divisive version, the algorithm starts off with the initial data set as one big cluster and then further partitions the big group into different smaller clusters using distance measures. The agglomerative version starts by looking at each single element in the data set and linking the two elements with the lowest distance measure. Both methods continue until they reach either the single element or the total set of items .

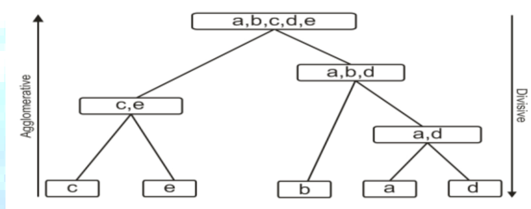

Fig-4: Hierarchical Clustering

3.9Sampling

Sampling is a powerful data reduction technique that has been applied to a variety of data mining algorithms for reducing computational overhead. In the context of association rules, sampling can be utilized to gather quick preliminary rules. This may help the user to direct the data mining process by refining the criterion for "interesting" rules. Sampling can speed up the mining process by more than an order of magnitude by reducing I/O costs and drastically shrinking the number of transaction to be considered. The validity of the sample is determined by two characteristics the size of the sample and the quality of the sample.

The quality, in the context of statistical sampling techniques, refers to whether the sample captures the characteristics of the database. The highest quality sample would be an exact miniature of the database; it would preserve the distributions of individual variables and the relationships among variables [6].The quality of the sample for association rule mining can be improved by considering factors like transaction length and transaction frequency [7].

A number of studies were conducted to propose efficient methods for mining association rules by reducing either the CPU computation time or the disk access overhead .Some studies considered the usage of sampling techniques for reducing the processing overhead [12,13]. Most of the prior works on sampling have concentrated on speeding up the phase by running a frequent item set mining algorithm only on a small sample of the database . Chiefly, researchers have evaluated the viability of using sampling [6] to reduce the dataset size. While such methods have shown quite a lot of promise it has been observed by several researchers [13,6] that it is often very difficult to quantify, apriori, the quality of the results obtained for a given sample size [13], necessitating novel and more effective sampling-based association rule mining algorithms to foster better mining results.

3.10Hash-based item set counting

A hash technique is very efficient in generating the candidate item sets, in particular for the large two-item sets, thus greatly improving the performance bottleneck of the entire process.

Soo et al [14] proposed Direct Hashing and Pruning [DHP] algorithm, an effective hash based technique for mining the association rules. This algorithm employs effective pruning techniques to progressively reduce the transaction database size. DHP utilizes a hashing technique to filter the ineffective candidate frequent 2 item sets. DHP also avoids database scans in some passes as to reduce the disk I/O cost involved.

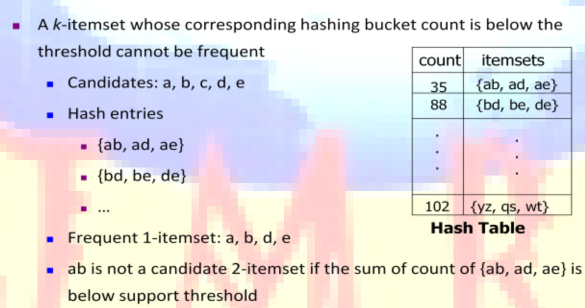

Fig-5: DHP- Reduce the Number of Candidates

Another novel hash-based approach [15] for mining frequent item-sets over data streams was developed by En et al. The algorithm compresses the information of all item sets into a structure with a fixed hash-based technique. This approach skillfully summarizes the information of the whole data stream by using a hash table to estimate the support counts of the non-frequent item sets, and keeps only the frequent item sets for speeding up the mining process.

Another algorithm Inverted Hashing and Pruning (IHP) proposed by John et al. It for is developed for mining association rules between words in text databases. The characteristics of text databases are quite different from those of retail transaction databases, and existing mining algorithms cannot handle text databases efficiently, because of the large number of item sets (i.e., words) that need to be counted. Two well-known mining algorithms, the Apriori algorithm [*1*] and Direct Hashing and Pruning (DHP) algorithm [*5*], are evaluated in the context of mining text databases, and are compared with the proposed IHP algorithm. It has been shown that the IHP algorithm has better performance for large text databases.

3.11Transaction Reduction

Transaction reduction is another way that helps in mining association rules effectively. It relies on a concept that a transaction that does not contain any frequent k-itemset is useless in subsequent scans.

AprioriTid algorithm [16] is another way of improving the performance of Association Rules. This algorithm is used to construct the frequent item set. The main idea of all these algorithm is according to the theory that the subset of a frequent itemed is a frequent item set and the superset of an infrequent item set is an infrequent item set. They scan the database repeatedly to mining the association rules. There is another feature for algorithm Apriori TID, the support of the candidate frequent item sets are calculated only at the first time it scanned the database D and also generated candidate transaction database D' which only includes the candidate frequent item sets. Then the latter mining are based on the database D', It reduce the time of I/O operation because D' is smaller than D, so, it enhance the efficiency of the algorithm.

3.12 Adding extra constraints

Another type of association rule mining involves in retrieving patterns by adding extra constraints on the structure of patterns. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

- Post-processing(filtering out patterns that do not satisfy user-specified constraints after the actual discovery process;
- Pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfy

pattern constraints.

- Dataset filtering(restricting the source data set to objects that can possibly contain patterns that satisfy pattern constraints)

Wojciechowski et al [10] proposed a constraint based algorithm that improves the efficiency of constraint based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset.

Rapid Association Rule mining (RARM) [18] is another method that uses a tree structure to represent original database and avoids candidate generation process. Constraints were applied during the mining process to generate only those association rules that are interesting to the users which guarantees the improvement of the efficiency of the existing mining algorithm.

Tien et al [19] presented a category based algorithm as well as the associated algorithm for constraint rule mining based on Apriori. This approach reduces computational complexity of mining process by passing most of the subsets of final itemsets.

3.13 Association Rule Clustering System.

Association Rule Clustering is useful when the user desires to segment the data. Lent et al [20] proposed a clustering Association rule in which they measure the quality of the segmentation generated by ARCS (Association Rule Clustering System) using the minimum description length (MDL) principle of encoding the clusters on several databases including noise and errors. Scale-up experiments show that ARCS, using the BitOp algorithm scales linearly with the amount of data.

Pi et al [4] proposed a new Fuzzy Clustering Algorithm on Association Rules for Knowledge Management. A fuzzy simulation degree and simulated matrix for association rules are put forward and a new algorithm based on dynamic tree is used for implementing the fuzzy clustering. The experimental results show that this algorithm clusters the Association rules efficiently.

Gupta et al [3] recently proposed a cluster based algorithm that uses a novel approach to the insignificant transactions dynamically. During a particular pass only those clusters that seem to be statistically useful are scanned and as a consequence all insignificant tuples are filtered dynamically. The results of the algorithm show that removing false frequent items and insignificant transactions dynamically improves the performance of association rule mining.

S.no	Algorithm	Technique	Memory utilization	Database	Time
1	Apriori	Generates Candidates using previous pass's large item-set. Uses join & prune step	Require more space and memory for candidate generation process	Suitable for Large Item-sets	Execution time is more for candidate generation process
2	Aprioritid	The *candidate* item-sets are generated the same way as in Apriori algorithm	Incurs addittional. cost if it can't fit into the memory	Suitable for small Item-sets	Execution time is better than apriori
3	Apriori hybrid	Apriori is used in the initial passes but we switch to AprioriTid in the later passes. The switch takes time, but it is still better in most cases.	Incurs addl.cost when shifting from apriori to apriori_tid	Suitable for Large Item-sets & small Item-sets	Execution time is better than apriori & apriori_tid
4	FP-growth	It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum support	Due to compact structure and no candidates generation requires less memory	Suitable for large and medium datasets	Execution time is large due to complex compact data structure
5	Partitioning	Partition the database for finding local frequent itemset first	Each partition is easily occupy in main memory	Suitable for large databases	Execution time is more because of finding locally frequent then globally frequent
6	Direct Hashing and Pruning	Use hashing technique for fining frequent itemsets	Require less space at earlier passes but more in later stages	Suitable for medium databases	Execution time is small for small databases

7	IHP	mining association rules between words in text databases	Require less space	Suitable for text databases	Has better performance for large text databases compared to DHP
8	Dynamic Itemset Counting	Based upon dynamic insertion of candidate items	Require different amount of memory at different point of time	Suitable for medium and low databases	Execution time is small because dynamic itemset are added according to situation
9	Vertical layout based techniqu e: Eclat	Use intersection of transaction ids list for generating candidate itemsets	Require less amount of memory compare to apriori if item sets are small in number	Suitable for medium and dense datasets but not suitable for small datasets	Execution time is small then apriori
10	Projecte d database techniqu e:H-mine	It uses the hyperlink pointers to store the partitioned projected database in main	Memory is utilized according to needs a partitions of projected database	Suitable for sparse and dense datasets	Execution time is large then FP-tree and others because of partition the database
11	Samplin g	Pick any random sample for checking frequency of whole database at lower threshold support	Very less amount of memory is needed	Suitable for any kind of dataset but mostly not give accurate results	Execution time is very much small

Table1: Comparison of various algorithms

Conclusion

Mining association rules is a prototypical problem as the data are being generated and stored every day in corporate computer database systems. To manage this knowledge, rules have to be pruned and grouped, so that only reasonable numbers of rules have to be inspected and analyzed. Thus an appropriate technique has to be employed to mine the association rules efficiently.

References

[1]. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conf., pp 487–499

[2]. B.Lent, A.Swami,J.Wisdom, "Clustering association rules", In the proc of 13th Int'l Conference on Data Engineering,pp.220.

[3]. Rajendra K.Gupta and Dev Prakash Agarwal,"Improving the performance of Association Rule Mining Algorithms by Filtering Insignificant Transactions dynamically", Asian Journal of Information Management, pp.7-17. 2009 Academic Journals Inc.

[4]. Pi Dechang and Qin Xiaolin," A New Fuzzy Clustering Algorithm on Association Rules for Knowledge Management", Information Technology Journal. Pp.119-124, 2008. Asian Network for Scientific Information.

[5]. Margaret H.Dunham,"Data mining Introductory and Advanced Topics", Pearson Education 2008.

[6]. Suzuki Kaoru, "Data Mining and the Case for Sampling," SAS Institute Best Practices Paper, SAS Institute, 1998.

[7]. Venkatesan T. Chakaravarthy, Vinayaka Pandit and Yogish Sabharwal, "Analysis of sampling techniques for association rule mining," In Proceedings of`the 12th International Conference on Database Theory, Vol. 361, pp. 276-283, 2009.

[8]. Kavya Naveen (PhD) Dr. M.V.Sathyanarayana N.C. NaveenData and Event Stream Mining,IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.4, April 2008

. [9]. Han, j. and Pei, J. 2000. Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter2, 2, 14-20.

[10]. Wang, C., Tjortjis, C., Prices: An Efficient Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 2447, 2002. pp. 77-83.

[11]. Yuan, Y., Huang, T., A Matrix Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 3664, Sep2005.pp 370-379.

[12]. Thevar., R.E; Krishnamoorthy, R," A new approach of modified transaction reduction algorithm for mining frequent item set", ICCIT 2008.11th conference on Computer and Information Technology.

[13]. Cheung, D., Han, J.Ng, V., Fu, A and Fu, Y. (1996), "A fast distributed algorithm for mining association rules" in Proc of 1996 Int'l Conference on Parallel and Distributed Information Systems'. Miami Beach, Florida, pp.31-44.

[14].Soo, J., Chen, M.S., and Yu, P.S., 1997, "Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules" IEEE Transactions On Knowledge and Data Engineering, Vol.No.5. pp. 813-825.

[15]. En Tzu Wang and Arbee L.P. ChenData," A Novel Hash-based approach for mining frequent item sets over data streams requiring less Memory space" Data Mining and Knowledge Discovery, Volume 19, Number 1, pp 132-172.

[16]. Rakesh Agarwal, Ramakrishna Srikant," Fast Algorithms for Mining Association Rules" 20th Intl Conference on VLDB, Santigo, Chile, Set.1994.

[17]. Cheung, D., Xaio, Y., Effect of data skewness in parallel mining of association rules, Lecture Notes in Computer Science, Volume 1394,Aug 1998,pages 48-60.

[18].Parthasarathy, S., Zaki, M.J.J., Ogihara, M., Parallel data mining for association rules on shared-memory systems, Knowledge and Information Systems: An International Journal,3(1):1-29,February 2001.

[19]. Wojciechowski, M., Zakrzewiez, M., Dataset filtering Techniques in Constraint based Frequent pattern Mining, Lecture Notes in Computer Science, Volume 2447, 2002, pp77-83.

[20] B.Lent, A.Swami,J.Wisdom, "Clustering association rules", In the proc of 13th Int'l Conference on Data Engineering,pp.220.

BIOGRAPHIES

W.Sarada is a research scholar in Computer Science at Rayalaseema University, Kurnool. Her interests include Data Mining, Computer Networks and Software Engineering.

B.Lakshmi Prasanna received her M.C.A degree in 2009 from JNTU, Kakinada and she is pursuing M.Tech (CSE) at JNTU, Kakinada. Her interests include Data mining, Computer Networks, Image Processing.