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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Efficient Algorithms for Mining High Utility Item sets from Transactional Databases

[1]Dr.R.A.Roseline, [2]Elizabeth Amalorpava Mary

Assistant Professor, PG & Research, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India[1]

Research Scholar, PG & Research, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India[2]

ABSTRACT: Utility mining plays a major role in many transactions oriented real world applications. The ecommerce application or e-shopping application attracts users by analyzing and providing them the most utilized item sets by other customers. Utility mining is a process of finding the profitable item sets among the set of candidate item sets in terms of most utilized items. However, mining and finding the most utilized item sets among the large volume of set of candidate item sets result in more time consumption. A number of candidate item sets that are available in the database requires more memory consumption and time for processing along with it requires more computational overhead. This problem is resolved in this work by introducing the tree based search paradigm namely UPGrowth and UPGrowth+ that is used to construct a utility tree by discovering only the high potential utility item sets from the transactional database. The tree based methodology improves the accuracy level of retrieving the most utilized potential utility item sets. The computation overhead of tree construction increases in the case of presence of most potential utility item sets. This is resolved in this research work by enhancing the UPGrowth tree construction by integrating the random hashing technique. This approach l allocates memory for every candidate item sets by finding the amount of memory that is required and calculates the node count utility along the route path to allocate the memory space required by them. Experimental tests conducted proves that the proposed approach provides efficient handling of potential utility item sets than the existing approach in terms of improved time complexity and space complexity.

KEYWORDS: High utility item sets, Potential growth, Tree based search paradigm, super market data set

I. INTRODUCTION

Data mining is an emerging field in the real world environment especially in the field of research and development area. Data mining is used to extract the information from the set of unstructured and structured data. One of the most important applications of data mining is the mining and extracting the most utilized item sets from the most used databases. This process provides convenient way for the users to mine and discover the large volume of item sets which are mostly utilized in nature. Discovery of most utilized item sets from the databases often require mining each candidate item sets present in the database.

There are various types of pattern mining approaches available in the real world environment. High utility mining is the most popular pattern mining approach that is frequently used by the transaction based application. This high utility mining approach is used to mine and extract the most profitable transaction patterns that are available in the database based on their frequency of transaction and customer preferences. Mining most utilized item sets from transactional database is a complex process due to the presence of closure down property where the multiple records present in the database have more interrelation with each other which cannot be satisfied. This limitation needs to be handled with more concern for supporting transaction of high utility item sets.

The patterns in transaction database need to be complete for achieving the efficient transaction. Missing values in the transaction may lead to performance degradation in the utility mining process. This process needs to be well handled for supporting the users with mining and providing them the most utilized patterns. This research work focuses on providing the well utilized item sets from the transactional database with the concern of minimum time complexity

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

and memory consumption. Well efficient handling of the transactional database is often required in case of presence of more patterns present in the transactional database. This problem needs to be well handled for the efficient construction and reconstruction of the patterns in terms of highly utilized items.

The main contribution of this work is given as follows: Large volume of records in the transactional database is mined for extracting the important high utility item sets. The main goal of this research work is to provide the efficient handling of large volume of transaction data base for the efficient mining of important high utility item sets by representing them in the tree format and sorting them in descending order. This mining approach is used to reduce the computation overhead and the time complexity in terms of extracting the most profitable item sets.

The organization of this research work is given as follows: In this section, the detailed introduction about the utility mining and its importance in the real world environment is discussed in a detailed manner. In section 2, detailed discussion about the various related works which has been conducted previously that focused on providing the efficient utility mining is discussed. In section 3, detailed discussion about the proposed research scenario is given in which time and memory constrained handling of the high utility item sets is given. In section 4, performance evaluation based on this work is given and discussed in the detailed manner. Finally in section 5, overall conclusion of this work and final result obtained in the proposed research work is given.

II. RELATED WORK

In this section, several research approaches that has been conducted related to pattern mining is discussed. Several approaches that concentrated on mining high utility and frequent utility item sets in the flexible manner are discussed in the efficient manner. The detailed discussion is given as follows:

Rakesh Agarwal et al [1] introduced a novel methodology for reducing the computation overhead that arises while handling large volume of data. He introduced the pruning methodology which eliminates the unwanted data items present in the database in case of presence of large volume of data. This approach is often used to eliminate the unwanted data items present in the database by classifying them into two classes. Those are important and unimportant data. This classification often leads to missing values which is resolved by introducing the buffering concept in which all the data classified as important would be stored.

Mohammed et al [2] introduced several ways that are used to establish the association rules present between the different candidate item sets. This process ensures the way of handling procedure that was used for the mining the most related patterns that resides in the database. This is ensured by introducing the association rule mining approach in terms of different functioning methodologies which assures that correlation resides between the different patterns present in the database. This association rule mining is established by using the novel approach called the clique generation approach. This approach leads to an efficient handling of data with the concern for user interaction.

Hen.J et al [3] defined a novel approach namely FP-Growth which is used to improve the association rule extraction used in the previous works in a considerable rate. This approach leads to efficient mining of set of rules that are associated with the different types of patterns in an accurate manner. This method is based on the tree structure that mines and extracts the patterns that are associated with the mining rules in terms of different candidate item sets.

Jiawei Han et al [4] introduced an improved mechanism which introduced an early methodology for predicting the number of repetitions that occur in the database in terms of different construction profile. This is used to assure efficient handling of large volumes of data in an efficient manner by avoiding the redundant data which were published earlier. This is ensured by integrating the fragmentation and partitioning technique with the methodology named FP construction which was introduced earlier.

Balázes Rácz [5] introduced a pattern growth algorithm which supports large volume of data which are generated dynamically. This approach is used to mine the database that contains large volumes of data in terms of different items. The tree constructed is updated dynamically by adding details of patterns which were used earlier dynamically.

Gosta Gahne et al [6] array based optimization approach stores the item sets that were generated currently in the array, thereby speeding up the item sets extraction than the previous approaches. This approach is used to mine

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

different approaches in terms of various logical improvements. This method assures the long delivery of products in terms of different logical operators.

Vaibhav et al [7] introduced a prefix tree based approach which overcomes the limitation that may arise due to the downward and upward closure property of a database. This approach is used to eliminate the limitation that occurs due to the presence of different transaction properties of the candidate item sets in their behavior. This is done by appending the prefix details with the candidate item sets by mining the large volume of data sets.

Ke-Chung Lin et al [8] developed a novel approach for mining the frequent candidate item sets that were preset in the database in terms of their most usage level. This is done by frequent pattern mining approach which mines and finds frequency pattern present in the database in terms of their minimum utilization threshold value. Minimum utility value should be set as a low value for predicting the large frequent items accurately.

A.B.M Rezbaul Islam et al [9] improved the frequent pattern mining approach which was introduced earlier by appending the association rule mining approach with it. This process is accomplished by finding the most novel approach namely most frequent item set extraction algorithm. This approach mines and derives the large volume of data in a considerable manner.

III. IMPROVED HIGH UTILITY ITEM SET EXTRACTION APPROACH WITH MEMORY BALANCING

High utility item set mining is the most essential task in the real world environment where an utility item set is used to indicate the candidate item sets with more profit. The profitable item sets are required to be extracted from the transactional database for providing a flexible and convenient way for the users to identify the highly recommended item sets. This is enabled by introducing a methodology called pattern mining based on which high utility sets can be extracted. However it suffers from time complexity and more memory consumption. This is resolved in this work by introducing the tree based utility mining mechanism. This approach is used to retrieve the most profitable item sets that reside in the transactional database. The high utility items sets are found by comparing it with the minimum utility threshold value which is found to be low in the proposed mechanism than the existing approach. And also the memory consumption for every candidate item sets is less. This overall architecture of this approach is given as follows:

From figure 1, one can predict the functioning procedure of the existing and the proposed research methodologies in terms of improved research scenario. The overall flow of this work is given as follow

- Find the promising and unpromising items in the transaction database
- Construct the tree in terms of promising items and the support count
- Mine the patterns based on UPgrowth+
- Allocate the memory for the candidate item sets based on R-Hash technique

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

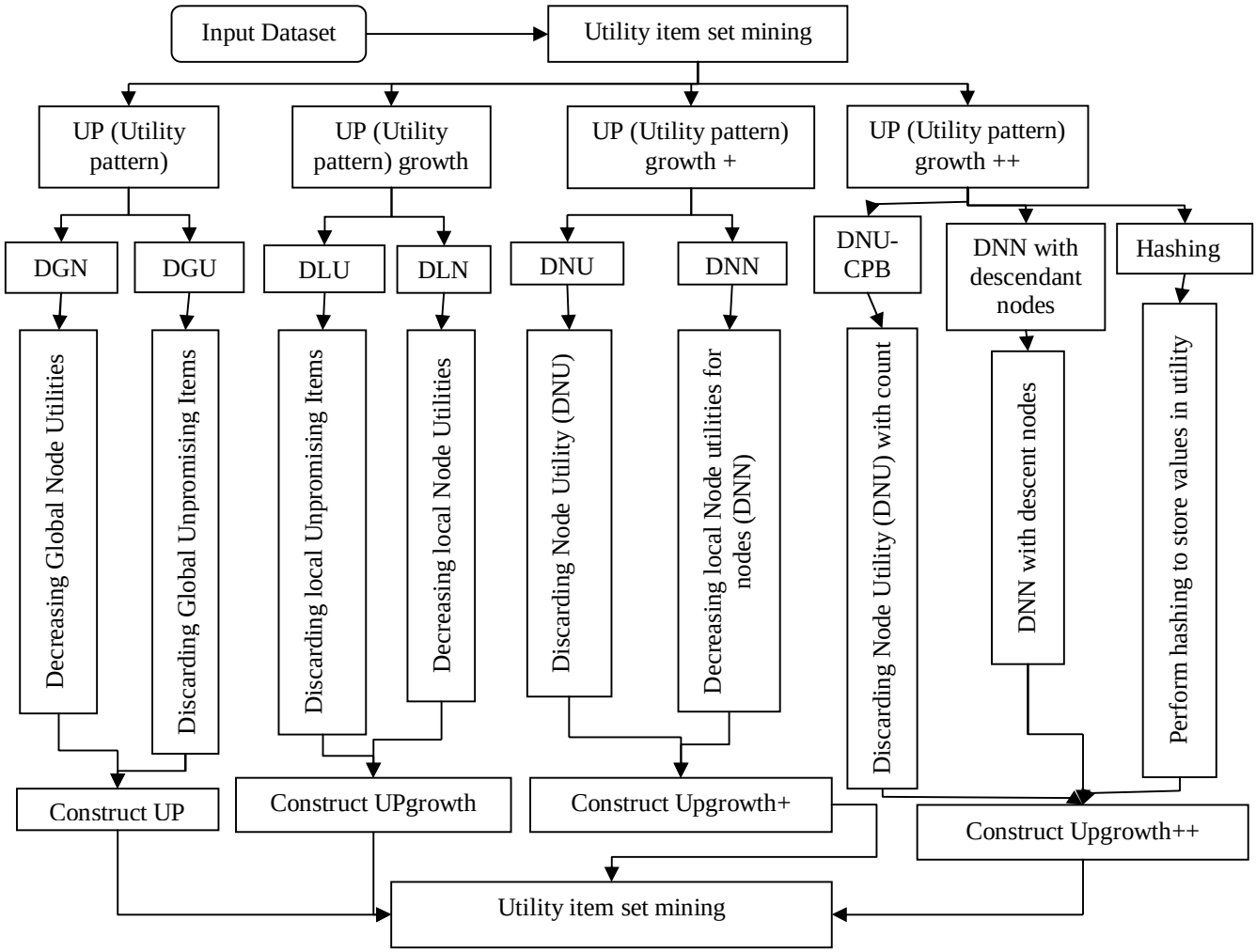

Figure 1. Architecture Diagram

These procedures are discussed in the proceeding sub sections.

A. *Finding the Promising and Unpromising Items in the Transaction Database:*

The transactional databases consist of a large number of transactional patterns in terms of different user transactions. High utility item sets are defined in terms of item sets in the data base that leads to a higher profit than the other candidate item sets. This high utility sets are found by comparing support count value with the available data sets in terms of different candidate item set values. The threshold values are set as all confident values are compared with the confidence value of the item sets that are available in the database.

Confidence value calculation procedure is given as follows: Consider a data base that contain set of items which is represented as I = {i1, i2,. ..in}. Transaction over a database is defined as the sub set of candidate item sets. That is transaction T is denoted as T contained in I. Each transaction is depicted with the identifier value in order to differentiate the other transaction patterns. The support of transaction pattern is defined as S(X) which in turn is defined as the number of time the corresponding pattern appears in the transaction database. The confidence value is calculated as follows:

$$all\_conf(X) = \frac{S(X)}{\max\{S(I_j)|\forall i_j \in X}$$

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

This calculated support value is compared with the min utility value based on which final retrieval is done. The patterns that are not having less support value than this calculated support value is considered for high utility mining. The candidate item sets that satisfied this condition is considered as the promising item and the other item sets are called as the unpromising item sets. The un promised item sets is identified by comparing the support value with the min utility threshold value. These candidate item sets which are defined as un-promised is eliminated from the transactional database to achieve efficient handling of the large volume candidate item sets.

B. *Constructing The Tree In Terms Of Promising Items And The Support Coun:*

Aim After eliminating the un-promised items from the candidate data sets, the efficient construction is done to make easy processing and retrieval of the high utility item sets in terms of their support value. This is done by sorting the promised candidate item sets in the descending order in order to know the high utility item set value. Every promised data set is retrieved from the list of set of sorted promised item sets. The extracted item set is fixed as the bottom node in the tree, from which paths are constructed to reach the root node. The root node construction is used to establish the valid path between the bottom node and the root node. The node utility value and the support count value is appended with every node in terms of different functional parameter values. The support count value is calculated for every node by subtracting the preceding nodes

The algorithm for reorganizing the path in terms of descendant values of minimum utility value is given as follows:

ALGORITHM: Insert_Reorganized_Path

1. If N has a child N_i, such that N_{ix}.item=i_x, increment N_{ix}.count by Pj.count. Otherwise, create a new child node N_{ix} with N_{ix}.item=i_x, N_{ix}.count=Pj.count, N_{ix}.parent=N and N_{ix}.nu=0.
2. Increase N_{ix}.nu
3. If there exists a node N_{ix} in Pj where x+1<m'
Call Inser_Reorganized_Path(N_{ix}, i_{x+1})

The above algorithm provides a convenient way for the users to sort and order the patterns that reside in the transactional database in terms of most utilized item sets. This will update the support count value of every node present in the environment by updating the values in the tree in terms of utility values of every preceding nodes.

C. Mining The Patterns Based On Upgrowth+

After constructing the Utility pattern tree, high utility mining is done by parsing the utility pattern tree from top to bottom. Thus the efficient and most utilized items from the transactional database with the consideration of priority are retrieved. This process enables a time and memory consumed mining of the high utility item sets from the transactional data sets. The working procedure of high utility set mining is given in the following algorithm

ALGORITHM: UPGROWTH+
Input: Up tree Tx, Header table Hx, item set X and min util threshold value in terms of different utility function values.
Output: High utility candidate item sets
1. For each entry i_k in H_x do
2. Trace each node related to i_k via i_k.hlink and accumulate i_k.nu to $nu_{sum}(i_k)$
3. If $nu_{sum}(i_k)$ >=min_util do
4. Generate a high utility item set Y = X U ik
5. Set pu(ik) as estimated utility of Y
6. Construct Y-CPB
7. Put local promising items in Y-CPB into Hy
8. Apply DLU to reduce path utilities of the paths
9. Apply Insert_Reorganized_path to insert paths into Ty with DLN

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

10. If Ty!=null then call UP-growth (Ty, Hy. Y)
11. End if
12. End for

The above algorithm is used to efficiently mine the high utility item sets from the transactional database in terms of minimum utility value.

D. Allocating Memory for the Candidate Item Sets Based On R-Hash Technique

In this section, memory handling problem of the proposed methodology is resolved by allocating the required memory for the high utility candidate item sets. This is done in three phases.

1. Scan the Data base to identify the promising item sets present in the database
2. Sort those item sets in the descending order based on their utility function value
3. Enhanced IFP is built to extract the relevant item sets

Random hashing technique assures the guaranteed allocation of unique memory space for the candidate item sets even in case of availability of more utility sets in the database. This provides secured framework even in case of malicious activities .

For any fixed set S of n keys, using a universal family guarantees the following properties.

1. For any fixed x in S, the expected number of keys in the bin h(x) is n/m. When implementing hash tables by chaining, this number is proportional to the expected running time of an operation involving the key x.
2. The expected number of pairs of keys x, y in S with x!=y that collide (h(x)=h(y)) is bounded above by n (n-1)/2m, which is of order $O(n^2/m)$. When the number of bins, m, is O(n), the expected number of collisions is O(n). When hashing into n^2 bins, there are no collisions at all with probability at least a half.
3. The expected number of keys in bins with at least t keys in them is bounded above by 2n/(t-2(n/m)+1). Thus, if the capacity of each bin is capped to three times the average size (t=3n/m), the total number of keys in overflowing bins is at most O(m). This only holds with a hash family whose collision probability is bounded above by 1/m. If a weaker definition is used, bounding it by O(1/m), this result is no longer true.

As the above guarantees hold for any fixed set S, they hold if the data set is chosen by an adversary. However, the adversary has to make this choice before (or independent of) the algorithm's random choice of a hash function. If the adversary can observe the random choice of the algorithm, randomness serves no purpose, and the situation is the same as deterministic hashing.

IV. EXPERIMENTAL RESULTS

Experimental tests were conducted by using synthetic dataset and it is compared against the existing approach to prove that the proposed methodology is better than its predecessor. Comparisons made against the parameters time and accuracy value are explained in the following sections.

A. Time Comparison

The time taken to mine the correlation pattern among both frequent and rate item sets are measured in existing and proposed methodology which proves that the proposed methodology provides better result than the existing approach

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Fig 2. Time comparison

The above graph proves that the proposed methodology provides the better result than the existing methodology by reducing the time consumption. In the x axis, methodology is taken and in y axis total time in milli seconds are taken.

B. Accuracy

In this section, accuracy value is measured which describes the accuracy of correlated pattern extraction. This accuracy value is measured for both the existing and proposed methodology and it is compared as in the following graph.

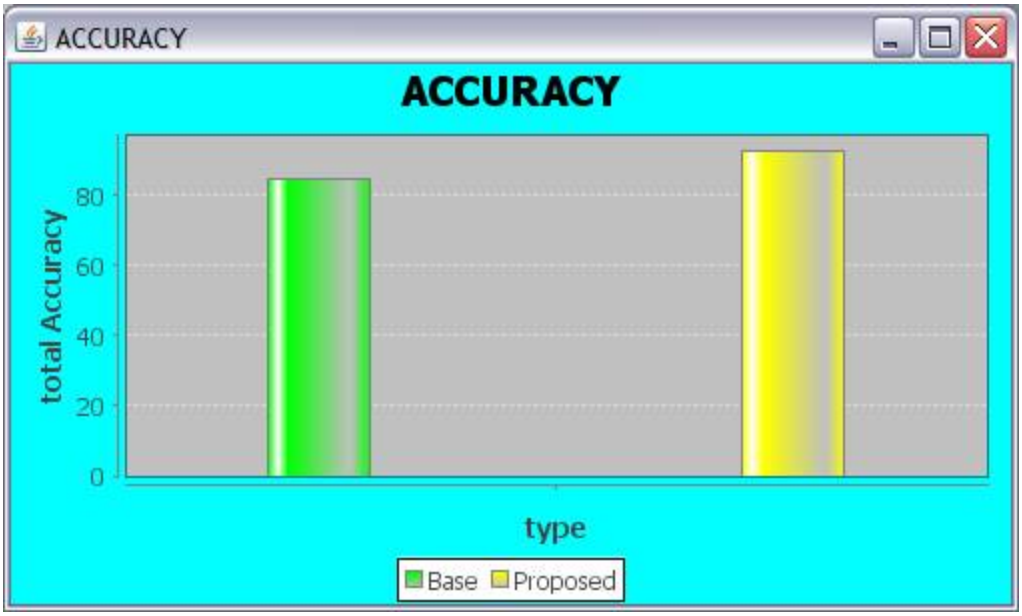

Fig 3. Accuracy Comparison

The above graph proves that the proposed methodology provides the better result than the existing methodology in terms of accuracy.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

V. CONCLUSION

Utility mining plays a greatest role in the real world environment and attempts to retrieve the most profitable terms present in the transaction database. The main problem in mining of larger transactional database for mining the high utility item sets are time complexity and memory consumption. This proposed approach leads to an efficient handling of these limitations by representing patterns in the tree format. Memory value is allocated for every candidate item sets individually by using the random hashing technique thereby avoiding the memory collusion problem. Experimental tests conducted prove that the proposed methodology provides better results than the existing research scenario in terms of improved accuracy.

REFERENCES

1. Agrawal, R, Imielinski, T, & Swami. A (1993), "Mining association rules between sets of items in large databases", In Proceedings of the ACM SIGMOD conference on management of data (pp. 207–216).
2. Zaki M.J, Parthasarathy. S, Ogihara. M, Li. W (1997)," New algorithms for fast discovery of association rules", In Proceedings of 3rd knowledge discovery and data mining conference (pp. 283–286).
3. Han.J, Pei.J & Yin.Y(2000), "Mining frequent patterns without candidate Generation", In Proceedings of the ACM-SIGMOD conference management of data (pp. 1–12).
4. Han.J, Pei.J & Yin.Y(2000), "Mining frequent patterns without candidate Generation", In Proceedings of the ACM-SIGMOD conference management of data (pp. 1–12).
5. Racz.B. (2004), "Nonordfp: An FP-growth variation without rebuilding the FP-tree", In Proceedings of IEEE ICDM workshop on frequent itemset mining implementations.
6. Grahne.G, & Zhu.J (2005)," Fast algorithms for frequent itemset mining using FP-trees", IEEE Transactions on Knowledge and Data Engineering, 17(10),1347–1362.
7. Vaibhav Kant Singh, Vijay Shah, Yogendra Kumar Jain, Anupam Shukla, A.S. Thoke, Vinay Kumar Singh, Chhaya Dule, Vivek Parganiha (2008), "Proposing an Efficient Method for Frequent Pattern Mining", World Academy of Science, Engineering and Technology.
8. Ke-Chung Lin, I-En Liao, Zhi-Sheng Chen (2011), "An improved frequent pattern growth method for mining association rules", Expert Systems with Applications, 5154-5161.
9. Rezbaul Islam.A.B.M & Tae-sun Chung (2011), "An improved frequent pattern tree based association rule mining technique", IEEE, 978-1-4244-9224-4.

BIOGRAPHY

Dr.R.A.Roseline is working as an Assistant Professor in the Department of Computer Science, Government Arts College at Coimbatore. She completed PhD in the area of wireless sensor networking at Bharathiar University. She completed M.Phil in the area of Networking at Periyar University. She completed MSC at Bharathiar University. She has published many national and International journals. Her research interests include Wireless sensor networking, mobile Adhoc network and cloud computing.

Ms. S.Elizabeth Amalorpava Mary received MCA degree from Karpagam College of Engineering, Coimbatore, India. She is pursuing Full Time M. Phil in the department of Computer Science under the guidance of Dr. R.A.Roseline at Government Arts College (Autonomous), affiliated to Bharathiar University, Coimbatore, India. Her area of research is Wireless Sensor Networking. Her research interests include Computer Networks, Information Security and Software Engineering.