IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011
ISSN (Online): 1694-0814
www.IJCSI.org

An Efficient Approach to Prune Mined Association Rules in Large Databases

D.Narmadha[1], G.NaveenSundar[2], S.Geetha[3]

[1]*Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India*

[2]*Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India*

[3]*Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India*

Abstract

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. However, when the number of association rules become large, it becomes less interesting to the user. It is crucial to help the decision-maker with an efficient postprocessing step in order to select interesting association rules throughout huge volumes of discovered rules. This motivates the need for association analysis. Thus, this paper presents a novel approach to prune mined association rules in large databases. Further, an analysis of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques are also discussed. We want to point out potential pitfalls as well as challenging issues need to be addressed by an association rule mining technique. We believe that the results of this approach will help decision maker for making important decisions.

Keywords- CLOSET, MAFIA, FP, Ontology, User constraint Template

1. Introduction

In **data mining**, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association rule Mining describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. An association rule is defined as the implication X => Y, described by two interestingness measures support and confidence where X and Y are the sets of items. Furthermore, valuable information is often represented by those rare-low support and discovered association rules are unexpected which are surprising to the user. As we increase the support threshold, the more efficient the algorithms are and the more the discovered rules are obvious, and hence, the less they are interesting for the user. As a result, it is necessary to bring the support threshold low enough in order to extract valuable information. Unfortunately, the lower the support is, the larger the volume of rules becomes, making it intractable for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules exceeds a limit. Thus, it is crucial to help the decision maker with an efficient technique for reducing the number of rules.

In this paper, a fairly comprehensive comparison of various association rule mining techniques is presented.

We analyzed the performance and the efficiency of different association mining approaches. Also this paper discusses the level of interestingness each technique provides. Rest of the paper is organized as follows. Section 2 contains a generalized summary of various association mining techniques and brief description of different approaches that have been taken for study. Section 3 gives a comparative analysis of various association mining techniques based on certain parameters. Section 4 discusses about the efficient approach to prune discovered association rules.

2. Methodology for Association Rule Generation

Association analysis has wide range of applications in market basket analysis, Intrusion detection, bioinformatics, web usage mining. There have been several such association mining techniques for generating association rules. Frequent item set concise representation as proposed by Burdick [2] and optimal rule discovery as proposed by Li [3], Zaki [4] introduced concise representation of frequent itemset to reduce the number of frequent itemsets and CLOSET algorithm was proposed [5] as a new efficient method for mining closed itemsets. Another solution for the reduction of the number of frequent itemsets is mining maximal frequent itemsets [6]. MAFIA algorithm is

based on depth-first traversal and several pruning methods. More recently, Bellandi [7] proposed ontology driven association rule extraction.The different approaches for the redundancy reduction of association rules are: Zaki and Hsiao used frequent closed itemsets in the CHARM algorithm [8] in order to generate all frequent closed itemsets. They used an itemset-tid set search tree and pursued with the aim of generating a small nonredundant rule set [9]. To this goal, the authors first found minimal generator for closed itemsets, and then, they generated nonredundant association rules using two closed itemsets.
Pasquier et al. [10] proposed the Close algorithm in order to extract association rules. Close algorithm is based on a new mining method: pruning of the closed set lattice (closed itemset lattice) in order to extract frequent closed itemsets. Association rules are generated starting from frequent itemsets generated from frequent closed itemsets.

From the above association rule mining techniques, few are selectively analyzed in detail in this literature.

2.1CLOSET: An Efficient Algorithm for mining Frequent closed Itemsets

This approach is an efficient algorithm for mining frequent itemsets with the development of three techniques:
(i) Applying compressed, frequent pattern tree FP-tree structure for mining closed itemsets without candidate generation.

(ii) Developing a single prefix path compression technique to identify frequent closed itemsets quickly.

(iii) Exploring a partition based projection mechanism for scalable mining in large databases.

Optimization1: Compress transactional and conditional databases using an FP-tree structure: FP-tree compresses databases for frequent itemset mining. An FP tree is a prefix tree structure representing compressed but complete information for a database. Its construction is simple. The transactions with same prefix share the portion of a path from the root. Similarly conditional FP tree can be constructed for conditional databases.

Optimization2: Extract items appearing in every transaction of conditional database: If there exists, a set of items Y appearing in every transaction of the X-conditional database, XUY forms a frequent closed item set if it is not a proper subset of some frequent closed item set with the same support.

Fig. 1 shows how the frequent closed item sets can be extracted directly from FP tree. This reduces the size of FP-tree because the conditional databases contain less number of items after extraction and also reduces the level of recursion.

Optimization3: Directly extract frequent item sets from FP-tree:

- This allows the program to identify frequent closed item sets quickly.
- Reduces the size of remaining FP tree to be examined.
- Reduces the level of recursion.

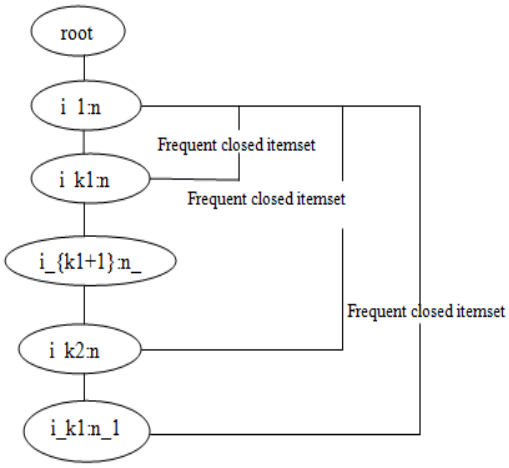

Fig. 1 Directly Extract frequent closed itemsets from FP tree

Optimization4: Prune search branches: Let X and Y are two frequent items with same support. If XCY and Y is closed itemset, there is no need to search for X conditional database because there is no hope to generate frequent item set from there. This reduces the overhead in searching for database.

2.2Ontology-Driven Association Rule Extraction

This provides an integrated framework for extracting constraint-based Multi-level Association Rules with an ontology support. This method can improve the quality of filtered rules.

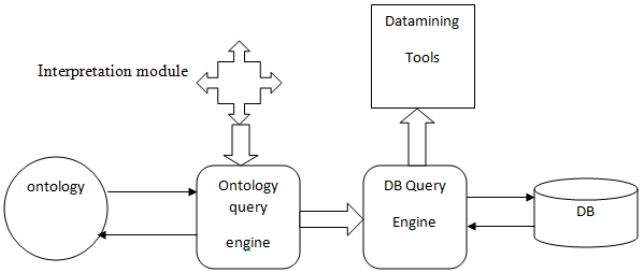

Fig. 2 System Architecture

The System Architecture as shown in Fig. 2 presents a view of set of components in the Ontology-Driven rule extraction. The ontology (OD) describes the domain of interest (D) and it is used as a means of meta-data representation. The interpretation module translates the requests of a user (user constraints) into a set of formal constrains (QD defined on OD) so that they can be

supplied to the Ontology Query Engine by means of a suitable query language. The aim of these constraints is to exclude some items from the output association rules, or to characterize interesting items according to an abstraction level. It includes both pruning constraints, used for filtering a set of non-interesting items, and abstraction constraints, which permit a generalization of an item to a concept of the ontology. By using pruning constraints, one can specify the exclusion of a set of items from the input transactions set, and, as a consequence, from the extracted rules.

There are several ways to reduce the computational complexity of Association Rule Mining and to increase the quality of the extracted rules: (i) reducing the search space; (ii) exploiting efficient data structures; (iii) adopting domain-specific constraints. The first two classes of optimizations are used for reducing the number of steps of the algorithm, for re-organizing the itemsets, for encoding the items, and for organizing the transactions in order to minimize the algorithm time complexity. The third class tries to overcome the lack of user data-exploration by handling domain-specific constraints.

This paper focuses on these optimizations by representing a specific domain by means of ontology and driving the extraction of association rules by expressing constraints. The aim of this work is to reduce the "search space" of the algorithm and to improve the significance of the association rules.

2.3 Selecting the Right Objective Measure for Association Analysis

This approach describe several key properties one should examine in order to select the right measure for a given application. An algorithm is presented for selecting a small set of patterns such that domain experts can find a measure that best fits their requirements by ranking this small set of patterns.Objective measures such as support, confidence, interest factor, correlation, and entropy are often used to evaluate the interestingness of association patterns. However, in many situations, these measures may provide conflicting information about the interestingness of a pattern. This approach describe several key properties one should examine in order to select the right measure for a given application.
The specific contributions are:
1) An overview of 21 objective measures is discussed in the statistics, social science, machine learning, and data mining literature. It is shown that application of different measures may lead to substantially differing orderings of patterns.
2) Several key properties are proposed that will help analysts to select the right measure for a given application. A comparative study of these properties is made using the twenty-one existing measures. Our results suggest that we can identify several groups of consistent measures having similar properties.
3) This also illustrates two situations in which most of the measures become consistent with each other, namely, when support-based pruning and a technique known as table standardization are used. This method also demonstrates the utility of support-based pruning in terms of eliminating uncorrelated and poorly correlated patterns.
4) An algorithm is used for selecting a small set of tables such that domain experts can determine the most suitable measure by looking at their rankings for this small set of tables.

2.4 An Approach to Facilitate the Analysis of the Association Rules

The goal of the research presented in this paper is to enable the end users to analyze, understand and use the extracted knowledge in an intelligent system or to support in the decision-making processes. In the paper, the GART algorithm is proposed, which uses taxonomies to generalize association rules, and the RulEE-GAR computational module, that enables the analysis of the generalized rules. This method uses iterative taxonomy to generalize and then prunes redundant rules at each step.

During years, manual methods had been used to convert data in knowledge. However the use of these methods has become expensive, time consuming, subjective and non-viable when applied at large databases. The problems with the manual methods stimulated the development of processes of automatic analysis, like the process of Knowledge Discovery in Databases or Data Mining.

In the Data Mining process, the use of the association rules technique may generate large quantities of patterns which make it difficult for the analyst to analyze the resultant pattern. A way to solve the problem of the large quantities of patterns extracted by the association rules technique is the use of taxonomies in the step of post-processing Knowledge. The taxonomies may be used to prune uninteresting and/or redundant rules. In this paper the GART algorithm and the RulEE-GAR computational module is proposed. The GART algorithm (Generalization of Association Rules using Taxonomies) uses taxonomies to generalize association rules. The RulEEGAR computational module uses the GART algorithm, to generalize association rules, and provides several means to analyze the generalized rules. The RulEE-GAR computational module that provides means to generalize association rules and also to analyze the generalized rules. The screen of the interface enables the user to analyze and to explore the generalized rules sets.

3. Analysis of Association Rule mining technique

Parameters used for Comparison

an increasing amount of attention during the last few years, and quite a number of theoretical results, algorithms and implementations have been presented that explicitly aim at improving the efficiency and Scalability of multi-relational data mining approaches. Table1 shows the comparison of different association rule mining approaches.

Table 1 Comparison Table

Parameters	2.1	2.2	2.3	2.4	Proposed Approach
Scalability	Yes	Yes	Yes	Yes	Yes
User Interesting criteria	No	No	No	No	Yes
Quality	No	Yes	No	No	Yes

Scalability: The system should be scalable with increase in amount of information.

User Interestingness Criteria: This depends on strong interaction with the user.

The comparison of different association rule mining approach is given in Table 1.
CLOSET is an efficient algorithm for mining frequent closed itemset.
Merits of CLOSET efficient algorithm:
1) Number of frequent items can be reduced.
2) Search space can be reduced.

DeMerits of CLOSET efficient algorithm:
This approach is based on statistical information and does not guarantee the rules are interesting for the user. There is no interactive approach to capture user interesting pattern.

Merits of Ontology Driven Rule extraction method:
The main advantages of the proposed framework can be summarized in terms of extensibility and flexibility.
1) The framework is extensible because data properties and concepts can be introduced in the ontology without either changing the relational database containing the transaction, or the implementation of our framework.
2) The flexibility is guaranteed from the separation of the data to analyze (the transactions) from the meta-data (description of the data).

The main parameters we considered for the analysis of different association rule mining approaches are scalability, quality of filtered rules, user interesting criteria. Efficiency and Scalability have always been important concerns in the field of data mining, and are even more so in the multi-relational context, which is inherently more complex.

Demerits of Ontology Driven Rule extraction method:
1) The overhead in conducting pruning tests and as a result the execution time is high.
2) This paper uses seRQL to express user knowledge which is not as flexible as rule schema.

Merits of Taxonomy in Association Analysis Method:
1) Efficient approach to prune and generalize association rules.
2) Good approach to analyze the rules generation.

Demerits of Taxonomy in Association Analysis Method:
1) This method uses iterative taxonomy in order to generalize and then prune redundant rules at each step which results in more number of iterations.

Merits and Demerits of objective measure selection method:
This paper describe several key properties one should examine in order to select the right measure for a given application. Objective measure is restricted only for data evaluation not sufficient to reduce number of rules and to capture interesting one.

4. Proposed Approach

An Efficient and Interactive Post mining of Association Rules (Proposed Approach)

The proposed approach is composed of two main phases.

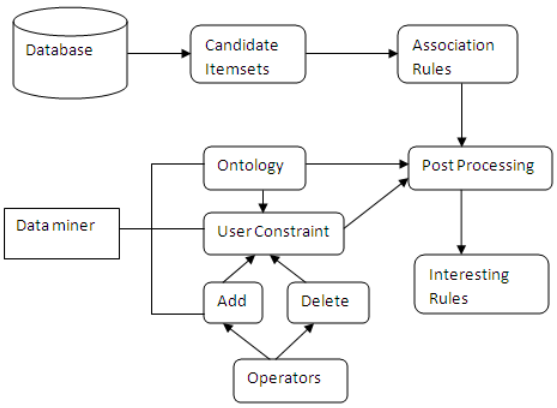

Fig. 3 Framework Description

The first phase includes the generation of support counts of item sets at each timeslot and candidate item

sets. The second phase involves mining of association rules from candidate items and post mining of association rules using ontology and user constraint template to guarantee user interesting rules as shown in Fig 3.

4.1Mining Candidate Item sets in large databases(Steps 1-3)

The transaction database is scanned using the lattice-dominant scan method which reads a whole transaction data set from time slot t_1 to time slot t_n for calculating the support count of each item. The cost of computing the support counts of all combinations of item sets at each timeslot increases with increase in the number of items. Hence, to reduce the computational cost, the interesting properties like tight upper and lower bounds of support counts of single items are used. It is then used to estimate support count of item sets at each timeslot without examining an input data set. The upper bound and lower bound support counts at each time slot indicates the range in which the true support count of an item set can be located. The lower bounding distance is used in the pruning of candidate

4.2. Post mining of Association Rules

As the number of attributes and the number of transactions becomes large, thousands of rules are from a database. As the number of rules become huge, it is difficult for the data miner to analyze the mining results. Also it is impossible to use the results. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. The interestingness of the rule strongly depends on interactivity with the user. Existing methods do not guarantee that interesting rules can be extracted. To select the interesting rule, the user knowledge should be expressed in an accurate and understandable form. In data mining, background knowledge ontology organizes domain knowledge and plays important roles at several levels of the knowledge discovery process. Ontology provides an explicit representation of concepts in a domain, where each concept is a collection if items. Instance of a concept represents the ground level items. The subsumption relation between concepts shows is-a super class, is-a subclass relations. The concept-instance relation represents the relation between concepts and the instances. There are two types of concepts: leaf-concepts and generalized concepts from the subsumption relation( ) (. Leaf-concepts are connected in the easiest way to database—each concept is associated to one item in the database. Generalized concepts are described as the concepts that subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts.

The Rule Schema formalism is based on the specification language for user knowledge introduced by Liu et al.The model proposed by Liu et al. is described using elements from an item taxonomy allowing an is-a organization of database attributes. Using item taxonomies has many advantages: the representation of user expectations is more general, and thus, filtered rules are more interesting for the user. However, taxonomy of items might not be enough. The user might want to use concepts that are more expressive and accurate. But, ontology includes the features of taxonomies and provides more representation power. In taxonomy, only the subsumption relationship is used to build the hierarchy. Thus, taxonomy is simply a hierarchical categorization or classification of items in a domain. In contrast, an ontology is a specification of several characteristics of a domain, defined using an open vocabulary. Dataminer develops ontology on the items in database. A user-constraint template is defined which allows the dataminer to select interesting rules according to his constraints. The user-constraint template can be represented in the following way:

UC<confectionery items=>grocery items>

We propose a matching operator for rule selection. The matching operator (M) selects the association rules that match with the user-specified constraint. When the matching operator is applied over user-constraint template M (UC), the antecedent and the consequent of the association rules should match.

1. Scan the transaction database to calculate the support count of the items at different timeslots (t1, t2, tn)
2. Lower Bounding distance is found between support count of each item (at various timeslots) and the minimum_support_sequence.
3. If lower bounding distance≤ user -specified threshold (U), the item set is considered as the candidate item set.
4. From the candidate item set, association rules are generated.
5. In the post processing step, ontology is constructed to describe the domain in which the analysis is done.
6. User-Constraint template is created to specify the interestingness of dataminer.
7. Addition and Deletion operator is applied over the user-constraint template to select the interesting rules.

Merits of proposed method:

1) This approach can prune and filter the discovered rules
2) Guarantees the rules are interesting for the user
3) The use of ontology provides specification of several characteristics of a domain

Demerits of proposed method:

1) The task of mapping ontology concepts with the DB items is time consuming

5. Experimental Results

The study is based on the supermarket dataset. The real dataset is in .arff (attribute relation file format).The dataset contains 217 attributes and 4627 instances. T (10000) shows the total number of transactions

Generating Association rules:

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given dataset.

In order to target the most interesting rules, we fix a minimum support of 2 percent, a maximum support of 30 percent, and a minimum confidence of 80 percent for the association rules mining process. Among available algorithms, we use the Apriori algorithm in order to extract association rules. The generated association rules describe the relationship between attribute.

Steps for Frequent item set and Association Rule Generation:

1. Scan the database to calculate the support of each item set.
2. Add the item set to frequent item set if support is greater than or equal to min_support.
3. At each level divide the frequent item set into left hand side and right hand side.
4. Calculate the confidence of each rule that is generated.
5. Generate strong rules satisfying min_support and min_confidence.

Fig 4. shows the user interface screen where the user can load the dataset. The dataset is in .arff file format. The input dataset contains header and data section.

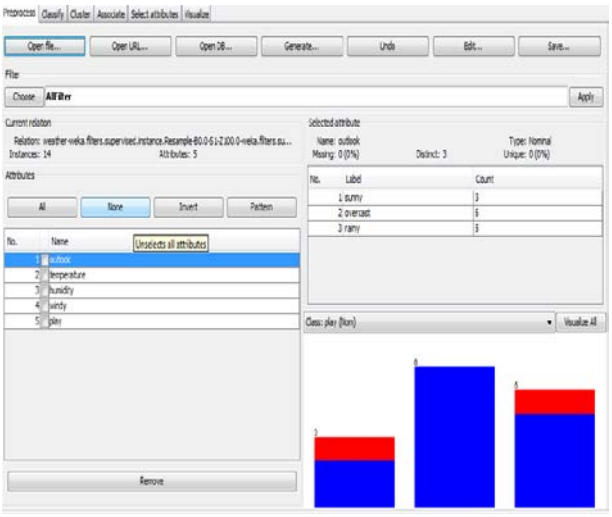

Fig. 4 Loading the Data set

Support(A=>B)=P(AUB) (1)

Confidence(A=>B)=P(B/A) (2)

Table 2 shows the comparison of number of interesting rules selected when matching operator is applied over user-constraint template and when not applied. The no. represents that each time different constraints are given to select different set of interesting rules. Our Experimental evaluation proves that the rules are generated and the selected rules are interesting to the user.

Table 2. Comparison of the Number of Rules with and without Applying Matching Operator

No.	Matching Operator	Without Operator
1.	225	1000
2.	162	1000
3.	289	1000
4.	77	1000
5.	139	1000

5. Conclusion

This paper discusses the problem of selecting interesting association rules through huge volumes of discovered rules. This motivates the need for association analysis

This paper discusses a novel efficient approach to prune mined association rules in large databases. A fairly comparative analysis of different association rule mining techniques for market basket analysis, highlighting strengths of different approaches, potential pitfalls as well as challenging issues need to be addressed by an association rule mining technique are also discussed. We believe that the results of this evaluation will help decision maker for making important decisions. We have evaluated the algorithms based on parameters like scalability, quality of filtered rules. Our evaluation shows that an efficient approach to prune mined association rules approach should be efficient and produce user interesting rules.

Acknowledgments

First and foremost, I praise and thank ALMIGHTY GOD whose blessings have bestowed in me the will power and confidence to carry out my work. I feel it a pleasure to be indebted to my guide, **Ms. S. Geetha, M.Tech,** Assistant Professor, Department of Computer

Science and Engineering for her invaluable support, advice and encouragement.
I also thank **Mr.G.NaveenSundar, M.Tech (Ph.D),** Assistant Professor, Department of Computer Science and Engineering for his valuable support in completing this work successfully

References

[1] Claudia Marinica and Fabrice Guillet, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," *IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 6, June 2010.*

[2] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "Mafia: A Maximal Frequent Itemset Algorithm," IEEE Trans.Knowledge and Data Eng., vol. 17, no. 11, pp. 1490-1504, Nov.2005.

[3] J. Li, "On Optimal Rule Discovery," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 460-471, Apr. 2006.

[4] M.J. Zaki and M. Ogihara, "Theoretical Foundations Of Association Rules," Proc. Workshop Research Issues in Data Mining and Knowledge Discovery(DMKD '98), pp. 1-8, June 1998.

[5] M.A.Domingues and S.A. Rezende, "Using Taxonomies to Facilitate The Analysis of the Association Rules," *Proc. Second Int'l Workshop Knowledge Discovery and Ontologies, held with ECML PKDD*, pp. 59-66, 2005.

[6] J. Pei, J. Han, and R. Mao, "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 21-30, 2000.

[7] A. Bellandi, B. Furletti, V. Grossi, and A. Romei, "Ontology- Driven Association Rule Extraction: A Case Study," *Proc. Workshop Context and Ontologies: Representation and Reasoning*, pp. 1-10, 2007.

[8] M.J. Zaki and C.J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemset Mining," Proc. Second SIAM Int'l Conf. Data Mining, pp. 34-43, 2002.

[9] M.J. Zaki, "Generating Non-Redundant Association Rules," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 34-43, 2005

[10] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient Mining of Association Rules Using Closed Itemset Lattices," Information Systems, vol. 24, pp. 25-46, 1999.

[11] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Objective Measure for Association Analysis," *Information Systems*, vol. 29, pp. 293-313, 2004.

[12] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Trans. Knowledge and Data Eng.* vol. 8, no. 6, pp. 970-974, Dec. 1996.

[13] R.J. Bayardo, Jr., R. Agrawal, and D. Gunopulos, "Constraint- Based Rule Mining in Large, Dense Databases," *Proc. 15th Int'lConf. Data Eng. (ICDE '99)*, pp. 188-197, 1999.

[14] J. S. Park, M. Chen, and P. S. Yu. An effective hash based algorithm for mining association rules. In *ACM SIGMOD Intl. Conf Management of Data*, May 1995.

[15] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. InkeriVerkamo. Fast discovery of association rules. In U. Fayyad and et al, editors, *Advances in Knowledge Discovery and Data Mining*, pages307–328. AAAIPress, Menlo Park, CA, 1996.

Narmadha received the B.E degree in computer science and engineering from Francis Xavier Engineering College, Tirunelveli in 2006.She is currently working toward the MTech degree in Computer Science and Engineering of Karunya University. Her main research interests are Association Rule Mining and Web mining

Naveen Sundar received the B.E degree in computer Science and Engineering from C.S.I Institute of Technology, Thovalai in 2002.He received the MTech degree from Karunya University, Coimbatore in 2006. He is currently working toward the PhD degree. He is working as an assistant Professor in Computer Science department of Karunya University. His main research interests are Association Rule Mining, Databases, Web mining

S.Geetha received the M.Tech degree in Information Technology from Anna University, Coimbatore in 2009 and B.E degree in Computer Science and Engineering from SASTRA University formerly known as Shanmugha College of Engineering, Thanjavur in 2002. She is currently working as Assistant Professor in Dept. of Computer Science and Engineering, Karunya University, Coimbatore. Her areas of interests are in Data Mining and Network Security. She is a member of Computer Society of India (CSI).