A Study of Different Association Rule Mining Techniques

R. Z. Inamul Hussain
Research Scholar
Sri Chandrashekhendra Saraswathi Viswa maha Vidyalaya University,
Enathur,Kanchipuram-631561 India

S. K. Srivatsa, Ph.D.
Senior Professor
Prathyusha Institute of Technology and Management
Aranvoyalkuppam, Poonamallee,
Tiruvallur Road, Tiruvallur-602025

ABSTRACT
Association Rule Mining (ARM) is one of the major data mining methods used to mine hidden knowledge from databases that can be used by an organization's decision makers to increase overall profit. However, performing ARM needs frequent passes over the entire database. Clearly, for large database, the role of input/output overhead in scanning the database is very important. In this paper, we provide some fundamental concepts related to association rule mining and survey the record of existing association rule mining methods. Obviously, a single article cannot be a entire review of the entire algorithms, yet we wish that the references cited will cover up the major theoretical issues, guiding the researcher in motivating research information that have yet to be explored.

General Terms
Algorithms, Support, Confidence, Huge Datasets,

Keywords
Association rule, Data mining, Classification, Fuzzy, Association Rule Mining, Very large Dataset

1. INTRODUCTION [1]
The applications of computers, database technologies and computerized data collected works techniques require huge amount of data to be stored into databases. It, thus, becomes necessary to calculate this data and turn it into helpful knowledge. Data mining or Knowledge Discovery in Database (KDD) emerges as a result to the data examination problem. One of the data mining techniques that is used to determine interesting rules or relationships between attributes in databases is the Association rules. These rules facilitate in discovering knowledge at numerous conceptual levels, which, in turn, give a range of accepting, from general to specific, for the primary data. Mining association rules from huge data sets has been a determined topic in current research into knowledge discovery in databases. It has been noticed that the current advances in data warehousing and OLAP technology is a training to organize data at multiple levels of abstraction.

Therefore, the main focus of this learning is discovery of capable methods for association rule mining. There are various ways to explore efficient mining of association rules.. One may use the Apriori algorithm to inspect data items at multiple levels of abstraction under the similar minimum support and minimum confidence thresholds. Second choice is the application of dissimilar minimum support thresholds and probably different minimum confidence thresholds as well as mining associations at dissimilar levels of abstraction. This leads to mining attractive association rules , which may not only determine rules, but may also have high potential to find nontrivial, informative association rules because of its flexibility for focusing the attention to different sets of data and applying different thresholds at different levels.

When a single support threshold is used it allows many uninteresting rules to be generated together with the interesting ones if the threshold is rather low, but do not allows many interesting rules to be generated at low levels if the threshold is more high. Therefore, in their study, substantial methods have been made on how to know and remove the redundant rules across various levels.

The discovery of association rules becomes a very vital task in the process of data mining. The idea of identifying such rules is obtained from market basket analysis where the aim is to mine patterns telling the customer's purchase behavior. An important technique of data mining is association rules mining which studies the buying behaviors of customers and thus improves the quality of business decisions. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the databases.

2. ASSOCIATION RULE MINING [2, 3]
Association Rule Mining [2] is a technique that has been familiar in the relationship between the data that is how most association rule mining in a variety of ways. Association mining that finds dependencies between the values of an attribute was introduced and has emerged as a famous research region.

There are two vital basic methods [3] for association rules, support(s) and confidence(c). Since the database is huge and users worry about only those frequently purchased substance, regularly thresholds of support and confidence are determined by users to remove those rules that are not so interesting or needful. The two thresholds are known as minimal support and minimal confidence respectively.

Support(s) of an association rule is defined as the percentage/fraction of tuples that contain A and B to the total number of tuples in the database. Assume the support of an item is 0.1%, it means only 0.1% of the transaction purchased this item. Confidence is defined as the percentage/fraction of the number of transactions that contain A and B to the total number of records that contain A. Various measures have been introduced to explain the strength of the relationship between item sets A and B such as support, confidence and interest. The definitions of these measures, from a probabilistic model are as follow:

Support $(A \Rightarrow B) \Rightarrow P(A,B)$, or the percentage of transactions in the database that have both A and B

Confidence $(A \Rightarrow B) \Rightarrow P(A,B)/P(A)$, or the percentage of transactions having B in transactions those having A

Interest(A ⇒B) ⇒P(A , B)/P(A)P(B) denotes a test of statistical independence.

There are various techniques can be used to generate strong association rules among huge number of rules. They are as follow

3. APRIORI ALGORITHM [4]

The Apriori algorithm is one of the major vital algorithms for association rule mining because many of the other algorithms are based on this principle or extensions of it. This algorithm is based on Main memory.

The algorithm runs in two stages. i.e. frequent itemsets generation and association rule generation. The frequent itemsets generation is again a two stage procedure:

• Candidate itemsets (Ck) generation i.e. all probable combination of items those are possible candidates for frequent itemsets.

• Frequent itemsets (Fk) generation- support for all candidate itemsets are generated and those itemsets whose support is greater than the user-specified minimum support are qualified as the frequent itemsets

4. SCALABLE ASSOCIATION RULES [5]

Scalability means that as a system gets huge, its performance increases correspondingly. For data mining, scalability means that by considering advantage of parallel database management systems and additional CPUs, we can solve a large number of problems without needing to change our underlying data mining environment. An algorithm is proposed for detection of Scalable Association Rules from huge set of multidimensional quantitative datasets using k-means clustering technique based on the variety of the attributes in the rules and Equidepth partitioning using scale k-means for getting good association rules with high support and confidence. The discretization method is used to produce intervals of values for every one of the attributes in order to produce the association rules. The outcome of the proposed algorithm created association rules with high confidence and support in representing appropriate patterns between project attributes using the scalable k-means.

4.1 Merits of Scalable Association Rules

It doesn't allow to lead to significant statistics. It provides a correct evaluation of initial perceptions of the visualizing methodology.

5. SAMPLING ALGORITHM [4]

A variety of sampling algorithms for association rule mining has been projected. The algorithm selects a random sample form the database and then generates frequent itemsets in the sample that uses the support which is less than the user specified minimum support for the database. These frequent itemsets are represented by S. Then the algorithm calculates the negative border of these itemsets denoted by NBd (S).The set of itemsets that are candidate itemsets but did not satisfy the minimum support are called as negative border. Simply NBd (Fk) = Ck - Fk After that for each itemset X in S U NBd (S) it checks whether X is frequent itemset in entire database by scanning the database. If NBd (S) contains no frequent itemsets then all the frequent itemsets are found.

If NBd (S) has frequent itemsets then the algorithm generates a set of candidate itemsets C_G by increasing the negative border of S U NBd (S) until the negative border is blank. Now for each and every itemset X in C_G the algorithm scans the database for the next time. If all the frenquent ietmset is found in first scan then it is base case.If it requires two scan over the database then it is worst case.

6. MINING LEVEL CROSSING ASSOCIATION RULES [6]

Mining level-crossing association rule at multiple idea level may lead to the finding of mining strong association among at different stage of hierarchy. A top-down progressive deepening technique is designed for mining level-crossing association rules in huge transaction databases by extension of some active multiple-level association rule mining techniques

A method for mining "level-crossing" association rules is uses a hierarchy information encoded transaction table. This is based on the following thought. Initially, a data mining query is usually in related to only a part of the transaction database, such as *food* instead of all the datasets. It is useful to first collect the related set of data and then work recursively on the task-relevant set.

Second, encoding can be done during the collection of task-related data and thus there is no extra "encoding pass" needed. Third, an encoding string, which represents a position in a hierarchy, need less bits than the equivalent object identifier or bar-code.

6.1 Merits of Mining Level Crossing Association Rules

A top-down progressive deepening method is designed for mining level - crossing association rules, which improves the existing single and multilevel association rule mining algorithms and discover methods for sharing data structure and in-between results across level. Deriving a recent filtered transaction tables at each processing stage, this technique will do least processing work and produces minimum candidate sets.

7. ASSOCIATION RULE MINING ON MULTIPLE DATASETS [7]

The methods for association rule mining on multiple datasets are discussed here. Recently with technology and information systems enabling organization has a data-storage system, but the issue is that those with a huge data set, which is hard in the association rule mining, because it needs a computer with a high-performance to proceed, which was followed by a cost raised. There is a method to solve this problem is to distribute the data set to be processed by multiple computers, by the computer, it does not need a good performance to processing in association rule mining. However, it may have a disagreement with the association rules in the process of combining association rules from each machine, and association rules from multiple datasets may be inefficient compared to the association rules from only data set. So in the method of combining association rules needs a technique to help fix the issues mentioned above.

Researches related to association rule mining on multiple datasets have to appear very small. Probably, due to the association rule mining on multiple dataset that is hard process of combined association rules from distributed data, association rules with well-organized close to that of association rule mining from one datasets. The researchers appeared, there was an inefficient comparison clearly.

7.1 Merits of Association Rule Mining on Multiple Datasets

Association rule mining from huge dataset, require a computer with a very high-performance to process and high cost. There is a method to overcome this problem is to distribute datasets to be processed by multiple computers.

The process combined association rules from distribute datasets take the same association rules and checking the conflict of the association rules.

8. NON-PARALLEL ASSOCIATION RULE MINING [8]

This type of mining algorithm works in two steps. The first step is for frequent item set generation. Frequent item-sets are found from all possible item sets by the help of measure called support count (SUP) and a user defined parameter called minimum support. Support count of an item set is defined by the amount of records in the database that contains all the items of that set. If the number of minimum support is too high, then minimum number of rules may be generated. Similarly, if the value is too low, a large number of rules may be established. Finding better rules from them may be another issue. After finding the frequent item-sets in the first step, the second step generates the rules using another user-defined parameter called minimum confidence.

Another drawback of these algorithms is the encoding scheme where separate symbols are used for each and every possible value of an attribute. This encoding scheme may be appropriate for encoding the categorical valued attributes, but not for encoding the numerical valued attributes as they may have various values in every record. To avoid this situation, some ranges of values may be defined. For each range of values an item is defined. This approach is also not suitable for all situations.

The generated rule may have a huge number of attributes involved in the rule thereby making it hard to understand. If the generated rules are not understandable to the user, the user will never utilize them. Again, since more significance is given to those rules, satisfying number of records, these algorithms may take out some rules from the data that can be simply predicted by the user. It would have been better for the user, if the algorithms can generate some of those rules that are actually hidden inside the data. These algorithms do not give any significance towards the rare events, i.e., interesting rules.

8.1 Merits of Non-Parallel Association Rule mining

It measures the excellence of generated rule by making one estimation principle, i.e., confidence factor or predictive accuracy. This principle estimates the rule depending on the number of appearance of the rule in the complete database. More the number of occurrences better is the rule.

9. PARALLEL ASSOCIATION RULE MINING [8]

In this scheme the physical interconnection of processors nodes is mapped into a logical ring of processor nodes, so that each processors node has a right neighbor and left neighbor.

At the early stage each processor nodes finds a partial measure of fitness for every individuals (rules) in its local subpopulation, by accessing only its local dataset.

Then each and every processor transfers its complete local subpopulation of individuals, and the value of their partially computed fitness function, to its right neighbor. When processor node receives a subpopulation of individuals from its left neighbor, it does the following tasks: (*i*) it finds the partial fitness measure of the received individuals on its local dataset; (*ii*) it combines this partial fitness measure with the earlier one of the received individuals to produce a recent fitness measure; (*iii*) it sends the incoming individuals, with their updated partial fitness measure, to its right neighbor. This repetitive process is done until all individuals have passed through all the processors and returned to their original processors, with their ending fitness value accordingly computed. The aforesaid system is appropriate to all processors groups. This reduces inter-process communication problem.

9.1 Merits of Parallel Association Rule Mining

It can achieve considerable speed up with a limited constraint.

Further the number of rules generated from these models is provided. The result is similar to that obtained using sequential algorithms. This is due to the fact that the parallelism is obtained only in fitness computation level and rest of the operation is same as sequential one.

10. SPATIAL ASSOCIATION RULE MINING [9]

Researchers mostly focus in structures of spatial objects and spatial/or Non spatial relationships that have spatial predicates e.g. adjacent to, nearby, inside, close to, intersecting, etc Spatial association rules can correspond to object/predicate relationships having spatial predicate. Since huge amount of spatial data have been gathered in various applications, ranging from Remote Sensing to GIS, Computer Cartography, Environmental Assessment and Planning. Even though some efforts were made to merge spatial mining with Spatial Decision Support System but for the most part researchers for spatial database are using a well known data mining approach-Apriori based association rule mining. There are two major limitations in existing approaches; the biggest being, that in a typical Apriori based spatial association the similar records are essential to be scanned again and again to discover out the frequent sets. This becomes cumbersome, as spatial data is already known to be large in size. As far as *sparsedata* is concerned, an Apriori based spatial association rule may even be used but when there is *dense data* there were other approaches giving good performance. Researchers talk about only the positive spatial association rules; they did not consider the spatial negative association rules. Negative association rules are very helpful in some spatial issues and are capable of extracting some valuable and previously unknown concealed information.

10.1 Merits of Spatial Association Rule Mining

A recent and efficient Spatial DSS having all the necessary support technologies like OLAP, Specialized Analysis and Reporting along with Spatial Association Rule Mining technique is required for future planners and decision makers; recent and effective methods are needed to integrate the related Information Technologies to find out knowledge from huge spatial databases. It constructs a highly compact P-tree which is usually substantially smaller than the original database and thus save the cost of subsequent mining process.

This method also deals with the cases where there exist multiple concept hierarchies.

11. NEGATIVE ASSOCIATION RULES [10]

Classical association rules consider only items enumerated in transactions. Such rules are known as positive association rules. Negative association rules also consider the similar items, but in accumulation consider negated items (i.e. absent from transactions).

Negative association rules are helpful in market-basket analysis to find products that conflict with each other or products that complement each other. Mining negative association rules is a tedious task, due to the fact that there are important differences between positive and negative association rule mining. The researchers attack two key issues in negative association rule mining: (i) how to successfully search for interesting item sets, and (ii) how to efficiently identify negative association rules of interest.

To determine the nature (positive or negative) of the relationship, a correlation metric was used. A modern idea to mine strong negative rules was introduced. They combine positive frequent item sets with domain knowledge in the form of classification to mine negative associations. Though, their algorithm is difficult to generalize since it is domain dependant and requires a predefined taxonomy. Another modern algorithm for creaming both positive and negative association rules where designed. They include on top of the support-confidence framework another measure called *mininterest* for a good pruning of the frequent itemsets generated.

12. MINING DYNAMIC DATABASE USING INCREMENTAL ASSOCIATION RULE [11]

In dynamic databases, new transactions are added as time advances. This may commence new association rules and some active association rules would become worthless. Thus, the upholding of association rules for dynamic databases is an important problem. Maintaining association rules has been studied popularly in data mining. There are two main ways to mining association rules for dynamic databases incrementally. The initial approach assumes that association rules to be available are not stable over time. New patterns that represent latest trends to be revealed are more interesting than old patterns. Thus, the first approach treats current added transactions more important than old transactions. The association rules obtained from the first approach are just the rough calculation of those obtained by re-running Apriori algorithm.

13. COMBINING CLUSTERING AND ASSOCIATION RULE MINING [12]

A technique of analyzing links between binary attributes in a huge sparse data set was proposed recently. At first the variables are clustered to obtain homogeneous clusters of attributes. Association rules are then mined in each cluster. A numerous clustering methods and compared the resulting partitions are used. They generated their clusters based on hierarchical techniques which are separated into two groups: ascendant techniques based on an agglomerative algorithm and descendant techniques performed by a divisive algorithm.

Once the clusters have been generated by the different methods, association rules were created on the different clusters. While their technique did succeed in finding association rules that could not be discovered without clustering, the inherent fault was in the clustering algorithms that they employed. None of the techniques proposed offered a better solution to scenarios where huge items overlap across clusters.

A further drawback with some of the existing transaction clustering algorithms is that they depend on some form of domain specific knowledge, thus limiting their range of applicability. Executing several different clustering techniques and then generating rules based on each of the clusters generated becomes prohibitively expensive in certain situations. This is especially true when clustering is employed over a range of datasets from different domains.

13.1 Merits of Combining Clustering and Association Rule Mining

These are used for extracting hidden facts from a large repository of raw data. This technique supports the business organization for customer support and future extension in their business.

14. CONSTRAINTS BASED ASSOCIATION RULE MINING [13]

In order to increase the effectiveness of present mining algorithms, constraints were used during the mining process to produce only those association rules that are attractive to users instead of all the association rules. With this technique lots of costs of mining those rules that are turned out to be not interesting can be saved. Generally constraints are given by users, it can be knowledge based constraints, data constraints, dimensional constraints, interestingness constraints or rule formation constraints. Constraints based association rule mining is to discover all rules from a known data-set meeting all the user-specified constraints. Apriori and its variants only utilize two fundamental constraints: *minimal support* and *minimal confidence*. Though there are two points, one is some of the generated rules may be helpfulness or not informative to individual users; one more point is that with the constraints of minimal support and confidence those algorithms may overlook some interesting information that may not convince them.

14.1 Merits of Constraints based Association Rule Mining

It can handle the problem of a conjunction and/or disjunction of anti-monotone sub-constraint. It can utilize the properties of constraints to prune search area or save constraint checking. Therefore, it is more efficient than other techniques

15. FUZZY ASSOCIATION RULES [14, 15]

It use fuzzy logic to convert numerical data's to fuzzy data's, like "Income = High", thus maintaining the integrity of information conveyed by such numerical data's. On the other side, crisp association rules use sharp partitioning to convert numerical data's to binary ones like "Income= [1000K and above]", and can potentially introduce loss of information due to these pointed ranges. Fuzzy Apriori and its various variations are the only familiar fuzzy association rule mining (ARM) algorithms available nowadays. Like the crisp version of Apriori, fuzzy Apriori is a slow and inefficient algorithm for very huge datasets (in the order of thousands of transactions)[15]

The fuzzy set theory [14] is superior to the interval techniques because fuzzy sets give a soft transition among member and non-member of a set. In these techniques, each and every quantitative attributes is replaced by some other attributes that partition the range of the original one using the fuzzy theory. A column belongs to each of these partitions in the table containing transactions.

15.1 Merits of Fuzzy Association Rule Mining

Thus, fuzzy association rules will be useful to improve the flexibility for the users in building any decisions or developing the fuzzy systems. If mining procedure also generates a large number of rules, it will be of limited use because a human user does not have the capability to investigate these rules

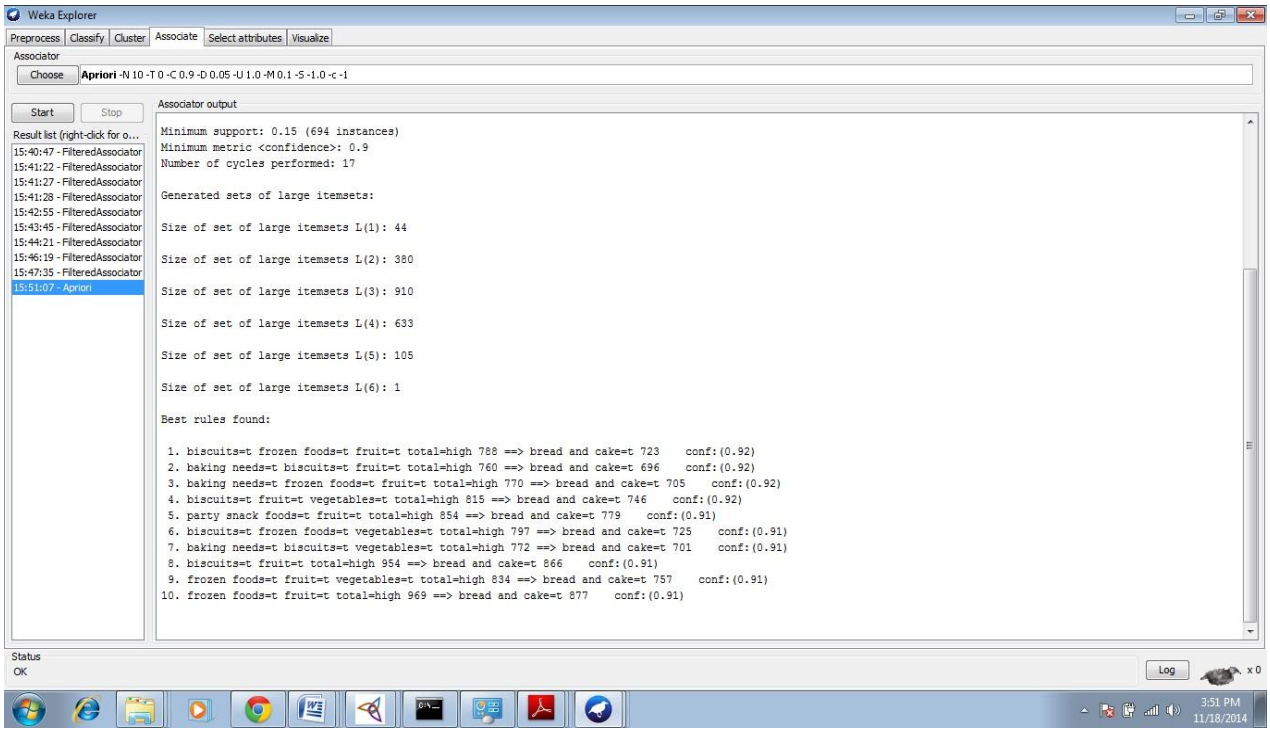

Fig 1: GUI of Association Rule Mining using WEKA tool

16. ASSOCIATION RULE MINING USING ONTOLOGY [16]

Association rules can be mined with the help of ontology. The system enables the definition of domain-specific constraints, by the help of the ontology to filter the instances used in the association rule mining process. This can improve the quality of the mined association rules and make them easier to understand.

The latter, which represents a powerful taxonomy, is used to explain the application domain by means of data properties. Explaining or updating these properties is a simple work and does not imply changing the items hierarchy, or the implementation level of our framework.

The ontology represents taxonomy powerful with data properties. A concept models a category of a collection of items, e.g. *bread* and *cheese* are kinds of *food*. Each data property explains an item feature at the considered level, e.g. *hasFlourType* is a feature of the products belonging to the *bread* concept only, while *hasPrice* is a feature of the product belonging to the *food* concept. It is vital to point out that the ontology is not the repository of data, but it gives the component for modelling the domain (i.e. the metadata). Adopting an ontology, the expert is able to focus the analysis only over a particular fragment of the data, without considering entire specific products. This enables to minimize the number of generated association rules and, usually, to improve the quality of the extracted model

16.1 Merits of Association Rule Mining using Ontology

The main advantages of this method can be summarized in terms of extensibility and elasticity: we can modify the algorithmic details or we can extend the ontology, without changing neither the implementation level nor the database transactions.

17. RESULTS

Association Rule mining can be done using various data mining tools and Fig 1 shows the output generated for Apriori algorithm using the sample dataset called supermarket using WEKA tool and the best rules are listed in the above figure.

18. CONCLUSION

Association rule mining has a lot of applications such market basket analysis, medical diagnosis, Website navigation examination, Native soil security and so on. In this paper, we surveyed the list of existing association rule mining methods and discussed about its advantages. We have given the research direction for Association rule mining techniques in which we have said about how to increase the efficiency and effectiveness of the Association rule mining and to get the interesting rules.

19. REFERENCES

[1] S. Srivastava et al, 2011 "On Performance Evaluation of Mining Algorithm for Multiple-Level Association Rules based on Scale-up Characteristics", Journal of Advances in Information Technology, VOL. 2, NO. 4.

[2] [Nuntawut et al., 2014] "A Technique to Association Rule Mining on Multiple Datasets", Journal of Advances in Information Technology, vol. 5, no. 2, may 2014.

[3] S. Kotsiantis, D. Kanellopoulos **"**Association Rules Mining: A Recent Overview**",** GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[4] P. Kandpal, *"* Association Rule Mining In Partitioned Databases: Performance Evaluation and Analysis",(Master Thesis) IIIT-Allahabad,India

[5] T. Siddiqui, M Afshar Aalam, and Sapna Jain, 2012 "Discovery of Scalable Association Rules from Large Set of Multidimensional Quantitative Datasets" journal of advances in information technology, vol. 3, no. 1

[6] R. S. Thakur *et al*., 2006 "Mining Level-Crossing Association Rules from Large Databases" Journal of Computer Science 2 (1): 76-81, 2006 ISSN 1549-3636

[7] N. Kaoungku et al, 2014 " A Technique to Association Rule Mining on Multiple Datasets" Journal of Advances in Information Technology, vol. 5, no. 2,

[8] [S. Dehuri, et al. 2006] "Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations" American Journal of Applied Sciences 3 (11): 2086-2095, 2006 ISSN 1546

[9] R. Vyas et al, 2007 "Exploring Spatial ARM (Spatial Association Rule Mining) for Geo-Decision Support System" Journal of Computer Science 3 (11): 882-886, 2007 ISSN 1549-3636

[10][R. Sumalatha et. al. 2010] "Mining Positive and Negative Association Rules", International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 2916-2920

[11] R. Amornchewin et al.2009 "Mining Dynamic Databases using Probability-Based Incremental Association Rule Discovery Algorithm", *Journal of Universal Computer Science, vol. 15, no. 12 (2009), 2409-2428*

[12] Yun Sing Koh, Russel Pears "Rare Association Rule Mining via Transaction Clustering", Conferences in Research and Practice in Information Technology (CRPIT), Vol.87, , Australian Computer Society,2008

[13] Anthony J.T. Lee et al, 2006 "Mining association rules with multi-dimensional constraints" The Journal of Systems and Software 79 (2006) 79–92

[14] [Z. Farzanyar et al. 2012] "efficient mining of fuzzy association rules from the pre-processed" Computing and Informatics, Vol. 31, 2012, 331–347

[15] A. Mangalampalli, V. Pudi, 2009"Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets" *IEEE International Conference on Fuzzy Systems*

[16] A. Bellandi et al, 2006 "Pushing Constraints in association rule mining: an ontology-based approach". IADIS International Conference WWW/Internet 2007 ISBN: 978-972-8924-44-7 © 2007 IADIS