ISSN:0975-9646

Association Rule Mining: A Survey

Gurneet Kaur

M.Tech Scholar, Department of Computer Science and Applications,
Kurukshetra University, Kurukshetra

***Abstract*: Association Rule Mining (ARM) has been the area of interest for many researchers for a long time and continues to be the same. It is one of the important tasks of data mining. It aims at discovering relationships among various items in the database. The objective of this paper is to present a review on the basic concepts of ARM technique along with the recent related work that has been done in this field. The paper also discusses the issues and challenges related to the field of association rule mining. A small comparison based on the performance of various algorithms of association rule mining has also been made in the paper.**

***Keywords*- Association rule mining, Apriori, Weka.**

I. INTRODUCTION

Data mining is the analysis step of the KDD(Knowledge Discovery and Data Mining) process. It is defined as the process of extracting interesting (non-trivial, implicit, previously unknown and useful) information or patterns from large information repositories such as: relational database, data warehouses etc. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining has been given much attention in database communities due to its wide applicability. The problem of mining association rules from transactional database was introduced in [1]. The concept aims to find frequent patterns, interesting correlations, associations among sets of items in the transaction databases or other data repositories. Association rules are being used widely in various areas such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc.[4]

Association rule are the statements that find the relationship between data in any database. Association rule has two parts "Antecedent" and "Consequent". For example {bread} => {eggs}. Here bread is the antecedent and egg is the consequent. Antecedent is the item that is found in the database, and consequent is the item that is found in combination with the first.

A more formal definition can be given as [7]: Let I = {i1,i2, …, in} be a set of items. Let *D* be a set of task relevant data transactions where each transaction *T* is a set of items such that $T \subseteq I$. A unique TID is associated with each transaction. Let *A* be a set of items. A transaction *T* is said to contain *A* if and only if $A \subseteq T$. An association rule is implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \text{null}$.

Association rule mining is done to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem of finding association rule is usually decomposed into two subproblems (see Figure 1) [18].

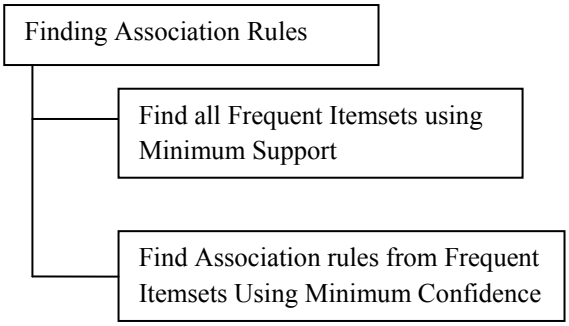

Figure 1: Generating Association Rules

As shown in figure 1 one sub problem is to find those itemsets whose occurrences exceed a predefined threshold in the database, those itemsets are called large or frequent itemsets. The second subproblem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Tk, Tk = {I1, I2, … , Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1} $\Rightarrow$ {Ik}, by checking the confidence this rule is determined as interesting or not. Then the remaining rules are generated by deleting the last items in the antecedent and inserting them to the consequent, thereafter the confidences of the new rules are checked to determine their interestingness. This process is repeated until the antecedent becomes empty. Since the second sub problem is quite simple, most of the researchers focus on the first sub problem.

The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation and frequent itemsets generation. The itemsets whose support exceed the support threshold are called as large or frequent itemsets and those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets. The two thresholds on which ARM technique is based are called as minimal support and minimal confidence respectively. Support is defined as the percentage of records that contain $A \cup B$ to the total number of records in the database. Let us assume the support of an item is 0.1%, it means only 0.1 percent of the transaction contain this item. Confidence of an association rule is defined as the fraction of the number of transactions that contain $A \cup B$ to the total number of records that contain A. Confidence is a measure of strength of the association rules, assume the confidence of the

association rule A B is 80%, it means that 80% of the transactions that contain A also contain B together [18].To illustrate this concept, a small example from the supermarket area has been used [7]. The set of items is *I={bread, egg, butter, cheese }*and a small database(Table I) containing the items (1 represents that item is present and 0 represents that item is not present in a transaction). An example rule for the supermarket could be *{bread, egg} => {butter}* meaning that if bread and egg are bought, customers also buy butter.

TABLE I: SAMPLE DATABASE FOR FINDING ASSOCIATION RULE

T	Bread	Egg	Butter	Cheese
T1	1	1	0	0
T2	1	1	1	0
T3	1	0	1	1
T4	0	1	1	0
T5	1	1	0	0

In the example database, the item set{bread, egg, butter} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).

The rule *{bread ,egg}=>{ butter}* has a confidence of 0.2/0.4 = 0.5, which means that for 50% of the transactions contain bread and egg (50% of the times a customer buys bread and egg, butter is bought as well).

I.I Generalised Association Rule Mining Algorithm

Many algorithms for generating association rules are presented over time. Some of the well known algorithms are Apriori, Fp-growth, AIS, Apriori-TID, Apriori Hybrid, Partitioning algorithms ,Tertius Apriori Algorithm and many more. Some of the parallel association rule mining algorithms based on Data and Task include CD(Count Distribution), PDM(Parallel Data Mining) , HPA(Hash-based parallel Mining of Association Rules) and PAR(Parallel Association Rules) and many more.

In general, a set of items (such as antecedent (LHS) or the consequent (RHS) of a rule) is called an itemset. The length of an itemset is given as the number of items contained in an itemset. Itemsets of some length k are called k-itemsets. Generally, an association rules mining algorithm contains the following steps[18]:

a) The set of candidate k-itemsets is generated by 1-extensions of the large (k-1)-itemsets generated in the previous iteration.
b) Support for the candidate k-itemsets are generated by a pass over the database.
c) Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until there are no more large itemsets in the database. The most commonly used approach for finding association rules is based on the Apriori algorithm. The efficiency of the level wise generation of frequent itemsets is improved by using the Apriori property which says that all nonempty subsets of a frequent itemset must also be frequent[11].

II. LITERATURE REVIEW

Various journals and articles concerning association rule mining algorithms were studied from year 2008 to 2013. Some compared association rule mining algorithms while some modified the existing algorithms to improve the performance.

Huaifeng Zhang et al [5] proposed an algorithm to discover combined association rules. Compared with the existing association rule, this combined association rule technique allows different users to perform actions directly. In their study, they have focussed on rule generation and interestingness measures in combined association rule mining. In combined association rule generation, the frequent itemsets among itemset groups are discovered to improve efficiency.

Pratima Gautam and K. R. Pardasani [12] presented an efficient version of Apriori algorithm for mining multi-level association rules in large databases to finding maximum frequent itemset at lower level of abstraction. They proposed a new, fast and an efficient algorithm(SC-BF Multilevel) with single scan of database for mining complete frequent itemsets. The proposed algorithm can derive the multiple-level association rules under different supports in simple and effective way.

Xunwei Zhou and Hong Bao [19] proposed an algorithm for double connective association rule mining for which a three table relational database is used. The rules are found among the primary keys of the two entity tables and the primary key of the binary relationship table.

Raja Tlili and Yahya Slimani [13] proposed a dynamic load balancing strategy for distributed association rule mining algorithms under a Grid computing environment. Experiments showed that the proposed strategy succeeded in achieving better use of the Grid architecture assuming load balancing and this for large sized datasets.

Anis Suhailis Abdul Kadir et al[2] provided the preliminaries of basic concepts of negative association rule and proposed an enhancement in Apriori algorithm for mining negative association rule from frequent absence and presence itemset. Relative interestingness measures were adopted to prove that the generated rules are also interesting and strong.

Guimei Liu et al [3] presented different methods to deal with the false positive errors in association rule mining. Three multiple testing correction approaches- the direct adjustment approach, the holdout approach and the permutation-based approach are used and extensive experiments have been conducted to analyze their performances. From the results obtained, all the three approaches control false positives effectively but among the three permutation–based approach has the highest power of detecting real association rules, but it is computationally expensive.

Somboon Anekritmongkol and M. L. Kasamsan [17] proposed a technique (Boolean Algebra Compress Technique) that will reduce the amount of time in reading data from the database. It has been found that through experiments that the time was reduced considerably.

Different authors have compared the performances of different association rule mining algorithms by

implementing them on various kinds of datasets. Jesmin Nahar et al [6] compared the various association rule algorithms on heart disease data predicting healthy and sick heart status. The three association algorithms used were Apriori, Predictive apriori and tertius algorithm. Based on the experimental results they concluded that Apriori algorithm is the best suited algorithm for this type of task. A similar work was done by Jyoti Arora et al [8] who performed a comparison of various association rule mining algorithms on Supermarket data and obtained the results using Weka data mining tool. The algorithms compared include Apriori association rule, Fp- growth and Tertius association rule. After comparing execution time by these three algorithms, author finds that Fp- growth is faster than other two algorithms.

Varoius authors have also tried to combine the association rule mining technique with either clustering or classification or both. Sunita B. Aher and Lobo L.M.R.J [9] combined the clustering (K-means algorithm), classification (ADTree classification algorithm) and association rule (Apriori algorithm) for course recommender system in E-learning and compared the results with using only association rule. The author finds that the combined approach is better than only Apriori as there is no need to preprocess the data. Ritu Ganda [14] performed an integration of clustering (K- Means algorithm) and association rule mining (Apriori) on kidney dataset using WEKA. The results show that integration gives more accurate and well defined rules in case of each cluster formed for kidney dataset.

III. ISSUES AND CHALLENGES

A lot of research work has been done in the field of association rule mining and various authors have proposed different algorithms in this field. Still there exist many issues and challenges in this field which need to be solved in order to get complete advantage of this technique. The main drawbacks of the association rule mining algorithms are[10]:

a) Obtaining non interesting rules
b) Huge number of discovered rules
c) Low algorithm performance

End users of association rule mining tools encounter several problems such as the algorithms do not always return the results in reasonable time. It is also found that the set of association rules can rapidly grow to be unwieldy, especially when we lower the frequency requirements.

Extracting all association rules from a database requires counting all possible combinations of attributes. Support and confidence factors can be used for obtaining interesting rules which have values for these factors greater than a threshold value. In most of the methods the confidence is determined once the relevant support for the rules is computed. However, when the number of attributes is large computational time increases exponentially. For a database of m records and n attributes, assuming binary encoding of attributes in a record, the enumeration of subset of attributes requires $m*2^n$ computational steps. For small value of n traditional algorithms are simple and efficient but for large values of n the computational analysis is infeasible[10].

The key element that makes association rule mining practical is the *minsup* i.e., the minimum support specified by the user. It is used to prune the uninteresting rules. But using only a single *minsup* means that all the items in the database are of the same nature. This may not be the case all the time. For example, in retailing business customers frequently buy those items which have less price while the items which have a higher price may not be bought too frequently. In such a situation, if the *minsup* is set too high, the generated rules will contain only those rules containing only those items which have low price and contribute less to the profit of the organization. On the other hand, if the *minsup* is set too less, many meaningless frequent patterns will be generated that will overload the decision makers. This type of situation is called as rare item problem[20].

Association rule mining has been very successful in various fields like commercial, social and human activities. But this technique poses a threat to privacy. One can easily disclose other's information by using this technique. So before releasing the database the sensitive information must be hidden from unauthorized access. It has been found that one of the current technical challenge in this field is the development of techniques that incorporate security and privacy issues. The association rule hiding problem aims at sanitizing the database in such a way that through association rule mining one will not be able to disclose the sensitive data and only the non-sensitive data will be mined[15].

IV. PERFORMANCE REVIEW

Many algorithms for generating association rules have been presented over time. Some of the well known algorithms are Apriori, Fp-growth, AIS, Apriori-TID, Apriori Hybrid, Partitioning algorithms, FP-growth Algorithm, Tertius Algorithm and many more. The advantages and disadvantages of some of the association rule mining algorithms are discussed in tabular form (Table II) :-

The AIS algorithm was the first algorithm to generate all large itemsets in a transaction database. The algorithm is used to find qualitative rules. This technique is limited to only item in the consequent. The AIS algorithm makes multiple passes over the database. The main problem of the AIS algorithm is that it generates too many candidates that later turn out to be small[1]. Another drawback of this algorithm is that the data structures required for maintaining large candidate itemsets are not specified. The Apriori algorithm developed by [1] is the most well known association rule algorithm. Apriori means "from what comes before" and uses breadth first search technique. Its implementation is easier than other algorithms and consumes less memory. However it has certain disadvantages also. It only explains the presence and absence of an item in transactional databases and requires a large number of database scan. Moreover the minimum support threshold used is uniform and the number of candidate itemsets produced is large.To overcome some of the bottlenecks of the Apriori algorithm Fp-growth algorithm was designed which is based on tree structure.

TABLE II: PERFORMANCE REVIEW OF SOME ALGORITHMS

Association Rule Mining Algorithm	Advantages	Disadvantages
AIS	1. An estimation is used in the algorithm to prune those candidate itemsets that have no hope to be large. 2. It is suitable for low cardinality sparse transaction database.	1. It is limited to only one item in the consequent. 2. Requires Multiple passes over the database. 3. Data structures required for maintaining large and candidate itemsets is not specified.
Apriori	1. This algorithm has least memory consumption. 2.Easy implementation. 3. It uses Apriori property for pruning therefore, itemsets left for further support checking remain less.	1. It requires many scans of database. 2. It allows only a single minimum support threshold. 3. It is favourable only for small database. 4. It explains only the presence or absence of an item in the database.
FP- growth	1. It is faster than other association rule mining algorithm. 2. It uses compressed representation of original database. 3. Repeated database scan is eliminated.	1. The memory consumption is more. 2. It cannot be used for interactive mining and incremental mining. 3. The resulting FP-Tree is not unique for the same logical database

The frequent itemsets are generated with only two passes over the database and without any candidate generation process thus making it faster than the Apriori algorithm. FP-growth uses a compressed representation of the database thus the irrelevant information are pruned. However it cannot be used for interactive and incremental mining system as changes in threshold value or new insertions in database may lead to a repetition of the whole process if we employ FP-tree method.

V. CONCLUSION

Association rules are widely used in various areas such as telecommunication networks, risk and market management, medical diagnosis, inventory control etc. This paper presents a review on association rule mining. Firstly a brief introduction about association rule mining is given which is the process of finding co-relations , frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories.A generalized association rule mining algorithm have been proposed. The paper surveys the research work done by various authors in this field. Some of the issues related to this field have also been presented which can help upcoming researchers to carry on their work. The advantages and disadvantages of some of the mining algorithms have also been presented in a tabular form.

ACKNOWLEDGMENT

I am thankful to Prof R.K Chauhan, Dept. Of Computer Science and Applications ,Kurukshetra University, Kurukshetra, India for his generous guidance and useful suggestions for this review work.

REFERENCES

1. Agrawal R., Imielinski, T., and Swami, " Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,1993.
2. Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar and Abdul Razak Hamdan, "Frequent Absence and Presence Itemset for Negative Association Rule Mining ", IEEE,2011.
3. Guimei Liu, Haojun Zhang and Limsoon Wong, "Controlling False Positives Iin Association Rule Mining" In Proceedings of the VLDB Endowment ACM,2011.
4. http://en.wikipedia.org/wiki/Data_mining
5. Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang, "Combined Association Rule Mining", PAKDD 2008, LNAI 5012, pp. 1069-1074, 2008 © Springer- Verlag Berlin Heidelberg 2008
6. Jesmin Nahar,Kevin S.Tickle,Shawkat Ali and YI-Ping Phoebe Chen, "Diagnosis Heart Disease using an Association Rule Discovery Approach" In Proceedings of the IASTED International Conference Computational Intelligence August 2009.
7. Ila Chandrakar and A. Mari Kirthima, "A Survey On Association Rule Mining Algorithms", In International Journal Of Mathematics and Computer Research, ISSN: 2320-7167, Vol 1, Issue 10, Page No. 270-272, November 2013.
8. Jyoti Arora, Sanjeev Rao and Shelza, "An Efficient ARM Technique for Information Retrieval In Data Mining " In International Journal of Engineering Research and Technology Vol 2,Issue 10, October 2013.
9. Lobo L.M.R.J and Sunita B.Aher, "Combination of Clustering, Classification and Association Rule based Approach for Course Recommender System in E-learning ", In International Journal of Computer Applications, Vol 39, February 2012.
10. Maria N.Moreno, Saddys Segrera and Vivian F. Lopez, "Association Rules: Problems, solutions and new applications" In Actas del III Taller Nacional de Mineria de Datos y Aprendizaje, pp. 317-323, TAMIDA 2005, ISBN: 84-9732-449-8 © 2005 Los autores, Thomson.
11. Nitin Gupta, Nitin Mangal, Kamal Tiwari and Pabitra Mitra, "Mining Quantitative Association Rules in Protein Sequences" Data Mining, LNAI 3755, pp. 273-281, 2006 © Springer- Verlag Berlin Heidelberg 2006.
12. Pratima Gautam and K.R. Pardasani, "Algorithm for Efficient Multilevel Association Rule Mining" In (IJCSE) International Journal on Computer Science and Engineering, Volume 02, No. 05, 1700-1704, 2010.
13. Raja Tlili and Yahya Slimani, "Executing Association Rule Mining Algorithm under a Gird Computing Environment" In PADTAD,July 2011.
14. Ritu Ganda, "Knowledge Discovery from Database using an Integration of Clustering and Association Rule Mining ", In International Journal Of Advanced Research in Computer Science and Software Engineering Vol 3,Issue 9,September 2013.

15. Sanjay Keer, Anju Singh, "Hiding Sensitive Association Rule Using Clusters of Sensitive Association Rule", International Journal of Computer Science and Network(IJCSN), ISSN: 2277-5420, Volume 1, Issue 3, June 2012.
16. Shweta and Kanwal Garg, "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes And Comparative Analysis Of Various Association Rule Algorithms" In International Journal Of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 6, June 2013.
17. Somboon Anekritmongkol and M. L. Kulthon Kasamsan , " The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model" In IEEE,2009
18. Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rule Mining: A Recent Overview" In GESTS International Transactions On Computer Science And Engineering, Vol. 32(1), pp. 71-82, 2006.
19. Xunwei Zhou and Hong Bao ," Mning Double-Connective Association Rules from Multiple Tables of Relational Databases " In IEEE,2008
20. Ya-Han Hu, Yen-Liang Chen, "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Timing Mechanism" © 2004 Elsevier B.V.