International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 02 (MAR-APR, 2016)

A Literature Review of Modern Association Rule Mining Techniques

Rupa Rajoriya, Prof. Kailash Patidar
Computer Science & engineering
SSSIST Sehore, India
rprajoriya21@gmail.com

***Abstract:*-Data mining is a technique that helps to extract important data from a large database. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. As a lot of information is gathered, with the quantity of information doubling each three years, data mining is becoming a vital tool to rework this information into data One common method to extract useful patterns from data is association rule mining. It is used in n number of applications. But the problem with the existing association rule mining methods is that they generate rules by taking more time and space. This becomes a real problem when rules are to be mined from a large data set.**

***Keyword:*-FP Tree, Association Rule, Warehouse, Mining. KDD Processes.**

I. INTRODUCTION

In data mining, association rule learning is a widespread and well researched technique for locating interesting relations between variables in massive databases. It is meant to spot robust rules discovered in databases using different measures of power. Based on the conception of robust rules, [1] introduced association rules for locating regularities between merchandise in large-scale dealing knowledge noted by point of sale (POS) systems in supermarkets. For example, the rule → found within the sales knowledge of a food market would indicate that if a client buys onions and potatoes along, he or she is probably going to additionally get hamburger meat. Such information is often used because the basis for decisions regarding marketing activity likes e.g. promotional evaluation or product placements. In addition to the above example from market basket analysis association rules are used these days in several application areas as well as web usage mining, bioinformatics and intrusion detection. As against sequence mining, association rule learning generally doesn't take into account the order of things either inside a transaction or across transactions.

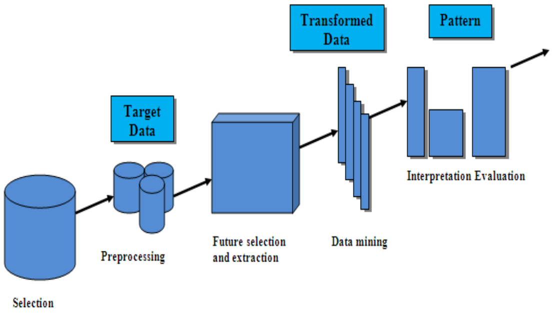


Fig.1 Concept of Mining Association Rule

Actual process work as follows. First we need to clean and integrate the databases. Since the data source may come from different databases, which may have some inconsistence and duplications, we must clean the data source by removing those noises or make some compromises. Suppose we have two different databases, different words are used to refer the same thing in their schema. When we try to unite the two sources we can only choose one of them, if we know that they indicate the same thing. And also real world data tend to be incomplete and noisy due to the manual input mistakes. The integrated data sources can be stored in a database, data warehouse or other repositories. As not all the data in the database are related to our mining task, the second process is to select task related data from the integrated resources and transform them into a format that is ready to be mined. Data mining is generally thought of as the process of finding hidden, non-trivial and previously unknown information in large collection of data.

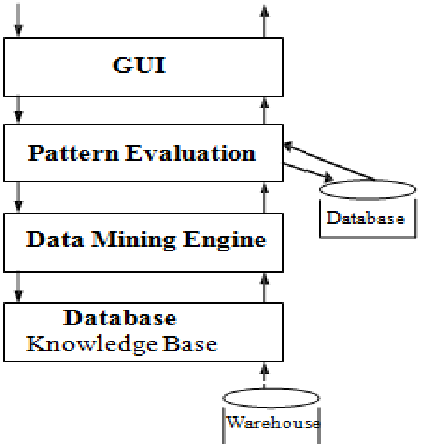


Fig. 2 Knowledge Discovery in Database processes

Association rule mining is an essential component of data mining. Basic objective of finding association rules is to find all co-occurrence relationship called associations. Most of the research efforts in the scope of association rules have been oriented to simplify the rule set and to improve performance of algorithm. But these are not the only problems that can be found and when rules are generated

II. RELATED WORK

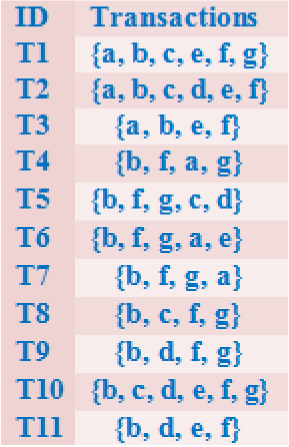
The chapter deals with relevant literatures review of various topics that fall under data mining and association rules. The first part discuses mining association rules definition and concept, and lastly some of the well-known data mining algorithms along with their computational difficulty.

III. ASSOCIATION RULE MINING

Association rule mining, one of the most important and well researched methods of data mining, introduced by [1]. Its objective is to extract interesting frequent patterns, correlations, associations or casual structures among sets of items in the transaction databases, data warehouse or other data repositories. Association mining is a fundamental data mining technique. It identifies items that are associated with one another in data. The problem of association rule mining is stated as follows. Let me= {a1, a2….an} be a finite set of items. A transaction database is a set of transactions T= {t1, t2...tm} where each transaction j⊂I (1≤ j≤ m) represents a set of items purchased by a customer at a given time. An itemset is a set of items A⊂I. The support of an itemset is denoted as sup (A) and is denoted as the number or transaction that contain A. An association rule A→ I is a relationship between two itemsets X, Y such that X, YI and X∩Y=Φ. In a database of transactions D with a set of n binary attributes (items) I, a rule is defined as an implication of the form X =>Y where X, Y are items and X∩Y =NULL the sets of items (for short item sets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively. Essentially, association mining is about discovering a set of rules that is shared among a large percentage of the data [11]. Association rules mining tend to produce a large amount of rules. The objective is to find the rules that are useful to users. There are two ways of computing usefulness, being subjectively and objectively. Objective measures involve statistical analysis of the data, such as support and confidence [1]. The Support, sup(X), of an item set X is defined as the proportion of transactions in the data set which contain the item set. Or support can be define as the rule X =>Y carries with support s if s% of transactions in D contain X U Y. Rules that have as greater than a user-specified support is said to have minimum support. The Confidence of a rule is defined as conf(X =>Y) = supp(X UY) / supp(X). Or confidence can be define as the rule X=>Y holds with confidence c if c% of the transactions in D that contain X also contain Y .Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

Association rule mining finds frequent patterns, correlations, associations or causal structures among sets of items or objects in transaction database, relational database, and other information repositories. The key step of association rule mining is to discover frequent itemset. Frequent patterns are pattern that occurs frequently in a database.

Table 3.1: Transaction Table



ID	Transactions
T1	{a, b, c, e, f, g}
T2	{a, b, c, d, e, f}
T3	{a, b, e, f}
T4	{b, f, a, g}
T5	{b, f, g, c, d}
T6	{b, f, g, a, e}
T7	{b, f, g, a}
T8	{b, c, f, g}
T9	{b, d, f, g}
T10	{b, c, d, e, f, g}
T11	{b, d, e, f}

The other definition of pattern defines it is a form, template, or model (or, more ideally, a set of rules) which can be used to make or to create things or parts of a thing. In data mining we can say that a pattern is a particular data behavior, arrangement or form that might be of a business interest. And an item set means a set of items or a group of elements that represents together a single entity. First we will introduce some naive and basic algorithms for association rule mining, Apriori series approaches. Then another milestone, tree structured approaches will be explained. Finally this section will end with some special issues of as-association rule mining, including multiple level ARM, multiple dimension ARM, constraint based ARM and incremental ARM.

A. AIS Algorithm focuses on improving the quality of databases together with necessary functionality to process decision support queries. In this algorithm only one item consequent association rules are generated, which means that the consequent of those rules only contain one item, for example we only generate rules like X ∩ Y -> Z but not those rules as X -> Y ∩ Z. The databases were scanned many times to get the frequent itemsets in AIS. The main drawback of the AIS algorithm is too many candidate itemsets that finally turned out to be small are generated, which requires more space and wastes much effort that turned out to be useless. At the same time this algorithm requires too many passes over the whole database.

B. Apriori Algorithm. Apriori is a great improvement in the history of association rule mining, Apriori algorithm was first proposed by Agrawal in [Agrawal and Srikant 1994]. The AIS is just a straightforward approach that requires many passes over the database, generating many candidate itemsets and storing counters of each candidate while most of them turn out to be not frequent. Apriori is more efficient during the candidate generation process for two reasons; Apriori employs a different candidate's generation method and a new pruning technique. Apriori algorithm still inherits the drawback of scanning the whole data bases many times. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements. Generally there were two approaches: one is to reduce the number of passes over the whole database or replacing the whole database with only part of it based on the current frequent itemsets, another approach is to explore different kinds of pruning techniques to make the number of candidate itemsets much smaller. Apriori-TID and Apriori-Hybrid [Agrawal and Srikant 1994], DHP [Park et al. 1995], SON [Savesere et al. 1995] are modifications of the Apriori algorithm. Most of the algorithms introduced above are based on the Apriori algorithm and try to improve the efficiency by making some modifications, such as reducing the number of passes over the database; reducing the size of the database to be scanned in every pass; pruning the candidates by different techniques and using sampling technique. However there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database.

C. FP-Treed (Frequent Pattern Tree) Algorithm. To break the two bottlenecks of Apriori series algorithms, some works of association rule mining using tree structure have been designed. FP-Tree [Han et al. 2000], frequent pattern mining, is another milestone in the development of association rule mining, which breaks the two bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. FP-Tree was introduced by Han et al in [Han et al. 2000]. By avoiding the candidate generation process and less passes over the database, FP-Tree is an order of magnitude faster than the Apriori algorithm. The frequent patterns generation process includes two sub processes: constructing the FT-Tree, and generating frequent patterns from the FP-Tree. The efficiency of FP-Tree algorithm account for three reasons, First the FP-Tree is a compressed representation of the original database because only those frequent items are used to construct the tree, other irrelevant information are pruned. Also by ordering the items according to their supports the overlapping parts appear only once with different support count. Secondly this algorithm only scans the database twice. The frequent patterns are generated by the FP-growth procedure, constructing the conditional FP-Tree which contain patterns with specified suffix patterns, frequent patterns can be easily generated as shown in above the example. Also the computation cost decreased dramatically. Thirdly, FP-Tree uses a divide and conquers method that considerably reduced the size of the subsequent conditional FP-Tree; longer frequent patterns are generated by adding a suffix to the shorter frequent patterns. In [Han et al. 2000] [Han and Pei 2000] there are examples to illustrate all the detail of this mining process.

IV. CONCLUSION

Association rule mining is a popular area of research. There are many real world applications are related to it. The problem with current algorithms is the duplicate rule generation. To overcome this problem we will proposed an updated algorithm.

REFERENCES

[1]. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207-216.

[2]. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 487-499.

[3]. Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In Eleventh International Conference on Data Engineering, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3-14.

[4]. Bayardo, R., Agrawal, R., and Gunopulos, D. 1999. Constraint-based rule mining in large, dense databases.

[5]. Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. 1997. Dynamic itemset counting and implication rules for market basket data. In SIGMOD 1997.

[6]. Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA, J. Peckham, Ed. ACM Press, 255-264.

[7]. Chen, M.-S., Han, J., and Yu, P. S. 1996. Data mining: an overview from a database perspective. IEEE Transaction on Knowledge and Data Engineering 8, 866-883.

[8]. Das, A., Ng, W.-K., and Woon, Y.-K. 2001. Rapid association rule mining. In Proceedings of the tenth international conference on Information and knowledge management. ACM Press, 474-481.

[9]. Garofalakis, M. N., Rastogi, R., and Shim, K. 1999. SPIRIT: Sequential pattern mining with regular expression constraints. In The VLDB Journal. 223-234.

[10]. Han, J. 1995. Mining knowledge at multiple concept levels. In CIKM. 19-24.

[11]. Han, J. and Fu, Y. 1995. Discovery of multiple-level association rules from large databases. In Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95), ZÄurich, Switzerland, September 1995. 420-431.

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

[13]. Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In 2000 ACM SIGMOD Intl. Conference on Management of Data, W. Chen, J. Naughton, and P. A. Bernstein, Eds. ACM Press, 1-12.

[14]. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I. 1994. Finding interesting rules from large sets of discovered association rules. In Third International Conference on Information and Knowledge Management (CIKM'94), N. R. Adam, B. K. Bhargava, and Y. Yesha, Eds. ACM Press,407.

[15]. Ng, R. T., Lakshmanan, L. V. S., Han, J., and Pang, A. 1998. Exploratory mining and pruning optimizations of constrained association’s rules. 13-24

[16]. Park, J. S., Chen, M.-S., and Yu, P. S. 1995. An elective hash based algorithm for mining association rules. In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, M. J. Carey and D. A. Schneider, Eds. San Jose, California, 175-186.

[17]. Pei, J. and Han, J. 2000. Can we push more constraints into frequent pattern mining? In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 350-354.