Volume 2, No. 3, May-June 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

# Study of Association Rule Mining Algorithms at Single Level of Abstraction

Shilpa Goel*
Department of Computer Science & Engineering
Haryana College of Technology & Management
Kaithal, Haryana, India
shilpa.goel12@gmail.com

Sunita Parashar
Associate Professor
Department of Information Technology
Haryana College of Technology & Management
Kaithal, Haryana, India
sunita.tu@gmail.com

Harvinder Singh
Assistant Professor
Department of Computer Science & Engineering
Haryana College of Technology & Management
Kaithal, Haryana, India
harvinderjabbal@gmail.com

***Abstract:*** Data mining helps in performing automated extraction and generating predictive information from large amount of data. The discovery of interesting association relationships among itemsets in large database which consist of transactions has been described as an important database mining problem and several algorithms for mining frequent pattern at single level have been developed. In this paper, we are reviewing different algorithms such as Apriori, FP-growth, Partition based algorithm, Incremental update (FUp based, probability based), Boolean Compress technique, Lattice Based Approach ,Fast algorithm to extract association rules from large itemsets.
***Keywords:*** Single-Level Association Rules, Data mining, support, Confidence

## I. INTRODUCTION

Today every organization is dealing with different data repository systems like relational databases, data warehouses, temporal databases, transactional databases, spatial databases, multimedia databases or WWW (World Wide Web) but no organization is taking advantage of their repositories. Because data to be stored is diverse in nature ranging from scientific, medical, demographic, financial to marketing as well as the volume of data is so much large that nobody has time to look at this data. Human attention has become the precious resource due to which every organization wants the relevant information to be taken care of. It is common for every company to analyze their databases to increase their profit. For this they must employ some technique to mine their databases, which is known as data mining. Data mining, *the extraction of hidden descriptive or predictive information from large databases*, is a powerful new technology with great potential to help companies and data analysts focus on the most important information in their data repositories. Data mining is a way to automatically analyze the data and to automatically classify it. This is one of the most active and exciting areas of the database research community [1]. Data mining refers to extraction or mining knowledge from large amount of data. The information and knowledge gain can be used for applications ranges from business management, production control and market analysis, to engineering design and science exploration. Data mining is the effort to understand, analyze, and eventually make use of the huge volume of data available. Through the extraction of knowledge in databases, large databases will serve as a rich, reliable source for knowledge generation and verification. This discovered knowledge can be applied to information management, query processing, decision-making, process control and many other applications. Figure 1 shows the control flow of data mining process.

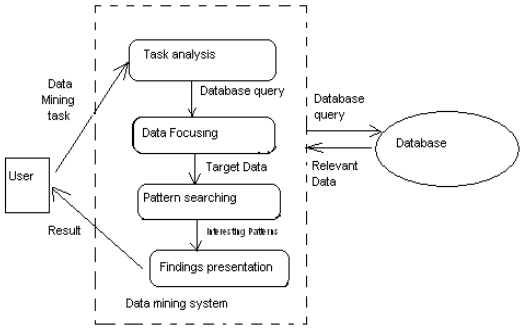


Figure 1 Control flow of data mining process

Knowledge Discovery and Data Mining have become areas of growing significance because of recent increasing demands for KDD (Knowledge Discovery in Data) techniques. There are several data mining techniques available to solve diverse data mining problems. They are mainly classified as associations, classifications, Summarization and clustering [1]. Association rule mining is an important data mining technique to generate correlation and association rule. The problem of mining association rules could be decomposed into two sub problems, the mining of frequent itemsets and the generation of association rules[3]. Frequent patterns are the patterns that appear in a database frequently (collection of set of items).

Finding such frequent pattern play an essential role in mining association, correlation, and many other interesting relationship among data. Thus frequent pattern mining is an important data-mining task and focused a lot in data mining research. Discovering frequent patterns is a very important data mining problem with a numerous of practical applications. As massive amount of data continuously being collected and stored, many industries are interested in mining such patterns from their datasets to increase their revenue. The discovery of frequent pattern helps in many business decisions making processes such as catalog design, cross marketing, customer shopping behavior etc. Frequent itemsets are used to generate association rules.

There are various association rule mining algorithms like single level mining association rule mining, generalized association rule mining and multilevel association rule mining**.** We will discuss only single level association rule mining**.** In single level association rules, one might find that 70 percent of customers that purchase bread may also purchase butter. This rule shows general information. The rules that are generated at single level shows strong associations.

Different single level mining algorithms are Apriori, FP-growth, Partition based algorithm, Incremental update (FUp based, probability based), Haskell based approach, Boolean compress technique, Lattice based approach, Fast algorithm. In this paper, we are studying all these algorithms for single level rule mining.

This paper is organized as follows. In Section 2, the concepts related to association rule mining are introduced. In Section 3, we are reviewing different algorithms for mining single level association rules. In section 4, we are discussing different approaches for mining. In section 5, we conclude the paper.

## II. ASSOCIATION RULES

Association rule mining is the discovery of associations or connections among objects. Association rule mining finds interesting associations among a large set of data items. Association rules identify the set of items that are most often purchased with another set of items. [2]

An Association rule is implications of the form X=>Y, Where $X,Y \subset I$, and $X \cap Y = \phi$. X is called the antecedent and Y is called the consequent of the rule. Each itemset has an associated measure of statistical significance called Support. For an itemset $X \subset I$ support (X) = s, if the fraction of transaction in database containing X equals s. A rule has a measure of its strength called Confidence that is defined as the ratio, support $(X \cup Y)$/support (X). For example, one may discover that a set of symptoms often occur together with another set of symptoms, and then further study the reasons behind this association [3]. Market basket analysis is just one form of association rule mining. In fact, there are many kinds of association rules.

Association rules can be classified in various ways[1], based on the following criteria:

**1**. Based on the types of values handled in the rule:

If a rule concerns associations between the presence or absence of items, it is a Boolean association rule. If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are partitioned into intervals.

**2**. Based on the dimensions of data involved in the rule:

If the items or attributes in an association rule references only one dimension, then it is a single dimensional association rule. If a rule references two or more dimensions, such as the dimensions buys, time of transaction, and customer category, then it is a multidimensional association rule.

**3.** Based on the levels of abstractions involved in the rule set:

Some methods for association rule mining can find rules at differing levels of abstraction. We refer to the rule set mined as consisting of multilevel association rules. If, instead, the rules within a given set do not reference items or attributes at different levels of abstraction, then the set contains single-level association rules.

**4**. Based on the nature of the association involved in the rule:

Association mining can be extended to correlation analysis, where the absence or presence of correlated items can be identified.

.

## III. ALGORITHMS FOR MINING SINGLE LEVEL ASSOCIATION RULES

Single level means that there are no hierarchies among items in an itemset. The goal of this algorithm is to find rules with high support and confidence. The single level association rules are most widely used to find out informative data. The single level association rules deals with the lowest level items of the concept hierarchy. The knowledge is said to be at a single level if the pattern involve only the raw data stored in database. Various algorithms are used to find single level association rules such as Apriori [2,3], FP-Growth [4]. The single level Association rule mining algorithms are broadly categorized as with and without candidate set generation algorithms.

### *A. Apriori Algorithm*

There are several approaches of mining single level association rules. The most popular and also the basis for many other algorithms is the Apriori algorithm[2,3]. It efficiently generates rules by reusing the data structure built during the determination of the support and the frequent itemsets by means of generating candidate itemsets. The data structure contains all support information and provides fast access for calculating rule confidences and other measures of interest. The problem with the raw data is that without good mining technique interesting information can't be predicted or found. Database systems do not provide necessary functionality for a user interested in taking advantage of this information. The problem that always appears during mining frequent relations is their exponential complexity because database is scanned n+1 number of times to generate large itemsets.

### *B. FP-Growth Algorithm*

FP-growth is used for mining the complete set of frequent patterns by pattern fragment growth. Efficiency of mining is achieved by compressing a large database into a condensed, smaller data structure, FP-tree which avoids costly repeated database scans [4]. This method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space. This need 2 scans of the database and candidate set generation will not occur in this algorithm. The only drawback is that resulting FP tree is not unique for same logical database.

### *C. Partition Based Algorithm*

Partition based algorithm divides the database into partitions that reduces the number of database scans to two[1]. During first scan some level-wise approach such as Apriori is used to find all large itemsets in each partition. During second scan, all large itemsets in each partition are used as candidates and counted to determine if they are large across entire database. If the items are uniformly distributed across partitions then a large fraction of itemsets will be large.

### *D. Fast Update algorithm*

FUP (Fast Update) algorithm provides incremental updation approach [5]. In this we separate winners (those that remain large in updated database) from loosers (that are not large in updated database) among large items in original database and find new winners that are large in original database (DB) and incremental database (db) i.e. (DB U db). This algorithm is 2 to 16 times faster than Apriori. When the increment is larger than the original size, the overhead decreases very rapidly from 10 to 5% but when the increment is much smaller than the original database, the overhead percentage ranges around 10 - 15% that is considerable**.**

### *E. Mining Frequent Patterns with Functional Programming*

This uses high-level declarative style of programming using a functional language (Haskell) for mining frequent patterns. In this, lattice is used as a data structure for search space of frequent pattern mining in which if superset is not frequent then all of its subset will not be frequent. Pattern matching (that is language feature commonly used with list as data structure) is done by means of Haskell[7]. A pattern is a set of items co-occurrence across a dataset. For a given candidate pattern, the task of pattern matching is to search for its frequency looking for the patterns that are frequent enough. The outcome of this search is frequent itemsets that suggest strong co-occurrence relationships between items in dataset.

### *F. Probability based Incremental Approach*

Probability-based incremental association rule discovery algorithm is used to extract interesting information from dynamic databases. This uses principle of Bernoulli trial to find expected frequent itemsets that reduces number of scans to original database[8]. As database updates it calculate 1-frequent and estimated frequent itemsets. By self joining 1-frequent itemsets of incremented database, we generate candidate-2 now separate new added C-2 that were not present in original database. These new C-2 are prune to generate Itemsets that will move to Temp Scan DB. Now calculate F-2, EF-2 of updated database (total database). The Itemset in Temp Scan DB need to be scan from Updated database that again reduce database scans.

### *G. Algorithm Based On Lattice Approac for Lower Cardinality Dense and Sparse Dataset*

The efficiency of Apriori is independent of cardinality, but the efficiency of proposed algorithm depends on the cardinality. In case of lower cardinality it provides better result for both sparse and dense dataset and good performance than Apriori, As the cardinality of the dataset increases performance of the proposed algorithm degrades specially for dense dataset. This algorithm does not follow generation of candidate-and-test method so it is free from join & prune process. This requires only two database scan first for 1-item and second for all the POSETs (partial ordered sets) [9]. This algorithm suffers from the discovery of the hidden relationship for higher cardinality dense dataset based on lattice approach.

### *H. Boolean Algebra Compress technique*

Boolean Algebra Compress technique compress data, reduce the amount of time to scan database by means of creating patterns of the transactions and reduce the size of file[10]. This algorithm is 10 times faster than apriori algorithm. Initially we find 1-large itemsets by means of support count. Then we compress data by means of creating patterns (pattern include itemsets that we get from L1 and are present in particular transactions). Now we generate k-large itemsets from compressed data instead of original database. This reduces number of database scans. The number of records does not affect the performance of Boolean Algebra Compress Technique for the Association Rule Mining but the number of itemsets affects the performance slightly.

## IV. DISCUSSION

We review different algorithms for mining. Apriori is based on candidate-set generation and test method. Contrary to this FP-growth is based on divide and conquer method where candidate set generation does not occur. In Functional programming, Haskell is used as programming language to implement Apriori algorithm. In FUp (Fast Update) incremental updation approach is applied on database to separate winners and loosers apart from this in probability based incremental approach bernoulii trial is used to find interesting information from dynamic databases. In Boolean Algebra Compress technique we compress data in the form of patterns that reduce database scans on the other hand in lattice based approach POSETS creation occurs that reduce the search space. Thus different algorithms are used to reduce database scans by means of reducing search space.

## V. CONCLUSION

In this paper, we dealt with algorithmic aspects of association rule mining. Mining Association Rules is one of the most used functions in data mining. Association rules are of interest to both database researchers and data mining users. We have provided a survey of previous research in this area as well as provide a brief comparison of different approaches to determine frequent itemsets and association rules between items of given transactions within the database. In future, for mining the association rules different researchers are working on different algorithms and data structures in order to refine the database. Data mining analysts are working in the direction to use lattice, hash table, tree based, list, array, pattern based as data structures along with mining algorithm to extract association rules.

## VI. REFERENCES

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques,” N. Harcourt India Private Limited ISBN:81-7867-023-2,2006.

[2] R.Agrawal, T.Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. of the ACM SIGMOD Conference on Management of data, Washington, D.C., May 1993, pp. 207-216.

[3] Rakesh Agrawal & Ramakrishan Srikant, "Fast algorithm for mining Association rules," IBM Almaden Research Center, 650 Harry road, San Jose, CA 95120: In proceedings of the 20th VLDB conference Santiago, Chile, 1994, pp. 487-499.

[4] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," In W.Chen, J. Naughton, and P. A.Bernstein, editors, 2000 ACM SIGMOD Intl. Conference on Management of Data, ACM Press, 05 2000, pp. 1-12.

[5] David W. Cheung, Jiawei Han, Vincent T. Ng, C.Y. Wong, " Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique," in proceedings of the 12th ICDE, New Orleans, Louisiania (IEEE),February 1996, pp. 106-114.

[6] M.H.Margahny and A.A.Mitwaly, "Fast Algorithm for Mining Association Rules," AIML 05 Conference, CICC, Cairo, Egypt,December 2005, pp. 19-21.

[7] Nittaya Kerdprasop, and Kittisak Kerdprasop, "Mining Frequent Patterns with Functional Programming," World Academy of Science, Engineering and Technology 2007.

[8] Ratchadaporn Amornchewin, Worapoj Kreesuradej, "Mining Dynamic Databases using Probability-Based Incremental Association Rule Discovery Algorithm," Journal of Universal Computer Science, Vol. 15, No.12, 28 June 2009, pp. 2409-2428.

[9] Ajay Acharya & Shweta Modi, "An Algorithm for Finding Frequent Itemset based on Lattice Approach for Lower Cardinality Dense and Sparse Dataset," International Journal on Computer Science & Engineering(IJCSE), ISSN:0975-3397, Vol.3, No. 1, January 2011.

[10] Somboon Anekritmongkol, Kulthon Kasamsan, "The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model," International Journal of Advancements in Computing Technology(IJACT), Vol. 3, No. 1, February 2011,pp.216-222.

 