International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 8 (September 2014) www.ijirae.com

Data Mining by Associated Rule in Systems

Venu Madhav Kuthadi[1] , Rajalakshmi Selvaraj[2]
[1]Department of AIS, University of Johannesburg, South Africa.
[2]Department of Computer Science, BIUST, Botswana

ABSTRACT-- The mining method by using the association rules gives the interesting association and co- relationships among the data sets. Very huge amount of data was collected and stored in many organizations and very much interested in mining rules combined with some sort of association in the databases. The discovery of relationships and association among the huge amount of data in business transaction helps in decisional making processes. The main features of association related are catalog design and cross marketing. Generally the association mechanism is related and notified in market basket analysis. The customer`s buying products and its association among the other products. The discovery of those relations and association helps the retailer to develop the marketing strategies and helps in gaining the insight into the items that are frequently purchased by the customer. The association of milk and bread will increase the sales of the retailer for the combination in a single visit. The mining of items with some form of association can be placed in proximity to encourage the sales of some items together. The paper is constituted the way of mining with interesting relationships among the items in the large data bases.

Keywords: Basket search, association rules, Boolean vector, threshold, item sets, maxpatterns, candidate set

1. INTRODUCTION

The mining of data items is to learn more about the items which are likely to retrieve frequently. It is performed on the basis of retail data of the transaction of the data base. The results may be used to plan the catalog design. The market basket plans helps the designer of the database in different store layouts. In many strategies the items which are frequently retrieved are placed together in a close proximity in order to further encourage the designer for mining operations easily. In many cases the customers who are eager to know about the specifications of the hardware will definitely purchase the soft wares. The combination of these items increases the speed of mining techniques in the data bases. In alternative strategies, if the hardware was placed at one part of the data base and the related software placing at another database may result in poor conditions of mining techniques. So, to avoid such strategies the association and the relationship among the data items is mandatory. If the database is consider as universe then the set of items available in the database, then each item has been given some sort of Boolean variables to represent the presence and absence of that item in the database. Every basket search is represented by Boolean variable as a vector of values that are assigned to these variables. The Boolean variable is analyzed for mining operations that reflects items which are frequently associated and retrieved together. The patterns can be represented in the form of an associated rule in the database.

2. ACCESS STRATEGY

The information retrieval mechanism is performed on the basis of association. The person who requires the information about the specification of the computer will surely knows about the software and it is represented in the form of association

Computer => software [software =10% and confident = 70%]

The association rule is designed with support and confident. The 10% of support describes the transaction the computer and the software are purchased together. The confident is 70% describes that many of the data sets are retrieved very easily with combinational rules from the database. The association rules that implemented in the data base have to contempt the threshold minimum support of 10% and threshold minimum confidence of 70%. These thresholds are set by the database designers.

Selections:

Let II= {i1, i2.i3….im} be the set of items. The DT is the task relevant data and the set of database transaction and each transaction TT is the set of items where TT ε II. every transaction is represented by an identifier Tid. The As is set of items. The transaction TT is said to contain As then As ε TT. The association rule can be an implication for like As=>Bs where As ε II and Bs ε II. The rule of As=>Bs is related to the transaction set DT that support s, where s is defined as the percentage of transaction in DT which contains both As and Bs. It is consider to be the probability of P(As U Bs). The rule of As=>Bs implications of confidence c in the transaction set DT and c is the percentage of transactions in DT which is having As and Bs. the conditional probability of P (Bs/As) is defined as

Supports (As=>Bs) = P (As U Bs).
Confidences (As=>Bs) = P (Bs/As)

The rules which satisfy the above support and confidence are as threshold minimum support and threshold minimum confidence. These two are combined and called as strong. The support and confidence value should occur in between the 0% to 100%. The itemset is a set of items containing the k items is the k itemset. The occurrence is related to the number of transactions containing the itemset. These are the measures of frequency and the count of the itemset. In many forms the itemset satisfies the minimum support of the occurrences of the itemset which is greater than the threshold minimum support and the total number of transactions of DT. The minimum support is satisfied by the number of transactions on the itemset that is referred to as the threshold minimum support count. If the itemset satisfied by the threshold minimum support then it is a frequent itemset. The association mining is based on two processes

1. First find the all frequent itemset and each of these itemsets will occur at least frequently as a predetermined threshold minimum support count.
2. The association rules are generated by using the frequent itemset and these rules must satisfy the threshold minimum support and threshold minimum confidence.

The overall performance is determined by the first step and the second step easiest of the two

3. MINING AS ROAD MAP

The rule based association between the absence and presentation of the items are considered as Boolean rule under association. The rules describe the association among the items and attributes and named as quantity based rule. The quantity rules for attributes are partitioned in the form of intervals. The quantity rule is described as

Person (XX1,"40…49) ∧ annual_income (XX1,''52K.…59K") => purchase (XX1, LCD TV)

The above association is considered as multi dimensions associated rule because there are three dimensions person, annual_income and purchase in multi dimensional rules there exists an abstraction based rule. It is described as

Person (XX1,"40…49) => purchase (XX1,"laptop")
Person (XX1,"40…49) => purchase (XX1,"computer")

The items are referenced at both levels of abstraction because the computer is the higher level of abstraction comparatively to the laptop. Such association rules are considered as multilevel rules of association. The association mining is a co-related mining such the presence and absence can be identified. The extension can be used for mining the frequent patterns closed items. Some patterns are correlated with their aliases. The aliases are useful in information retrieval because of its association with the original patterns. The association rules sometimes automatically tag the aliases improves the relational detection tasks. Some of the association orders determine the graph of relations between the patterns and its aliases. The association rules can also be extended for the maximum patterns which are the patterns occurrence in maximal frequency. The maxpattern are described as a frequent pattern, p1, such that it contains subpattern of p1 is not frequent. The subpattern is described as q1 is a subpattern of p1 and p1 is a subpattern of q1, that is, the q1 contains p1. The itemset can be closed frequent set where an item c1 is closed and if there is an superset of c1and c' Every transaction of c1 contains c'. Both the maxpattern and frequent itemset are used very much substantially to reduce the number of frequent item which are generated in the steps of mining operation. The itemsets recognized in such a way that which are frequent and local with respect to the entire data base, DT. The itemset that is potential with respect to the DT and frequent itemset in at least one of the partitions made in the data base. The local frequent itemsets are considered as candidate itemsets with respect to DT.

The mining techniques in the form of partitioning the data base is shown in the form of two phases in the figures

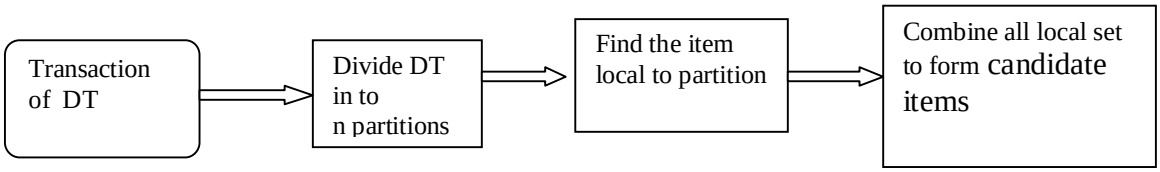

Fig.1 Phase.1 of DT

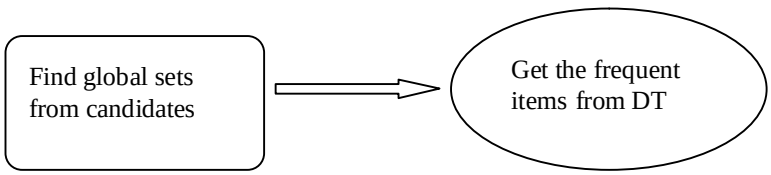

Fig.1 Phase 2 of DT

The frequent itemsets are collected from all the partitions and named as global sets of candidate with DT. In phase 2 the second scan of DT is performed so that the actual support of candidate is assessed to determine the global itemsets which are more frequent to occur. The partition size and the number of partitions are set in such a way that each partition can fit accurately in the memory in such a way that only read once in each phase.

4. FREQUENT PRUNING GROWTH

The test and generating methods reduce the size of candidate sets and may result to the good performance. The test and generation has to generate a very huge number of candidate sets. If there are 10^4 frequently generated 1-itemsets are available. The algorithm has to generate 10^7 candidates of 2-itemsets. The frequencies and test are accumulated in the candidate. To discover the frequent pattern of 100 the algorithm has to generate $2^{100} \supseteq 10^{30}$ candidate sets. For such type of operation it is recommended to scan the database repeatedly and check for the large set of candidates in the form of pattern matching. Thus, these operations are particularly meant for searching long patterns. The method of mining the complete set of items which occur frequently without candidate generations are simply frequent pruning growth (FPG). The FPG is simply a divide and conquer strategy which compresses the database that are representing frequent item sets in to frequent pruning growth and retain the association of the itemsets and divide the compressed databases in to some form of conditional databases. The conditional databases are meant as a specialized and projected databases and associated with the frequent item and the mining operation is performed on the database separately.

Algorithm

Input: list is denoted as L

Output: The nodes with prefix

1. First scan the database
2. Derive the set of frequent items and their frequencies.
3. Keep the minimum support 2 and sorted the order in descending support count.
4. Scan the database DT for second time and the items in each transactions are processed in L order
5. Scan the first transaction containing three items in the L order
6. Construct the first branch of the tree which the common node is the child.
7. The second transaction is linked to the root and share a common prefix with the existing path
8. Count each node along the common prefix.
9. The common prefix is incremented by 1.
10. Follow the prefix for the nodes.
11. The prefix are created and linked accordingly.

5. MINING OF FREQUENT PRUNING GROWTH

The mining of frequent pruning growth starts from each frequent pattern length and is consider as initial pattern. The pattern is constructed as conditional base pattern as a sub database which contains the set of all prefix paths in the frequent pruning growth tree. The co-occurrence of the suffix pattern construct the conditional FPG tree and performs recursively mining operations on the tree. The patterns increased on the basis of concatenation of the suffix pattern with the frequently occurrence of the patterns which are generated from the conditional FPG tree. The chart1 shows the how search method increase with respect to the quarter of searches in the data base.

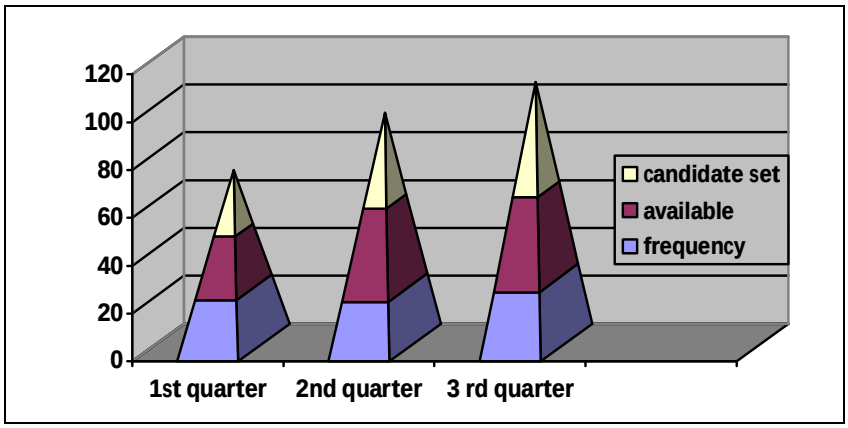

Fig.1 Frequency Search Method

The FPG increasing methods can see in the chart 1. The frequently search method increases the availability of the patterns in the database. The algorithm is repeatedly pruned and recursively operated on the database to find the availability of the percentage of the candidate patterns. The chart describes the problem of finding the frequent patterns which are looking for shorter and recursive patterns. The patterns are concatenating the suffix. The algorithm uses the least frequent items as a suffix based and gives very good selections and reduces the search costs.

6. CONCLUSION

When the database is large enough and sometimes it is very much unrealistic to build the main memory base FPG tree. The alternative is to partition the database into some form of projected databases. Now construction of FPG tree is performed and mining can be performed on the database. The process can be recursively performed to any type projected database even if the FPG tree is not fit in to the main memory. The study of the performance of the FPG tree shows that it is efficient method and very much scalable for both mining and short frequent patterns. The authors are requested to reduce theses recursively operations in the future course of study. The availability of the candidate set is very important for FPG tree. The candidate key is mainly base on the aliases and the availability of the pattern matches. It is one of the draw back because for one name there may be some many aliases. Sometimes the recursive procedure goes to the worst time if the main memory is not available. So, it is requested to reduce as much as, the recursive should be low and the availability of the candidate items should be high. In this paper the partition and selection of the patterns are based on the recursive procedures. It is requested to minimize the recursion techniques so the occupancy of the space in the main memory will decrease and the availability of the patterns increase in a simple search technique.

REFERENCES:

1. http//www.google.com for study materials.
2. Data Mining Concepts and Techniques second edition by Jiawei Han and Micheline Kamber pp 225-242
3. J. Han, M. Kamber, and A. K. H. Tung. Geographic Data Mining and Knowledge Discovery, chapter Spatial Clustering Methods in Data Mining: A Survey, pages 211-220. Taylor and Francis, 2001.
4. Data Mining by Margaret H.Dunham. Introductory and Advanced Topics Margaret H. Dunham.2002 Publisher: Prentice Hall pp 210-236
5. Modern data warehousing, mining and visualization by George M.Marakas
6. 6 Jul 2007 ... Download Free eBook: Data Mining: Concepts and Techniques,
7. Data Mining: Concepts and Techniques, 2 Edition ... August 26, 2010.
8. “Content based image retrieval with color space and texture features” proceeding of the 2009 an international conference on web information systems and mining.
9. Indexing and Mining One Billion Time Series icdm, pp.58-67, 2010 IEEE International Conference on Data Mining, 2010.
10. Discrete Mathematical structure by J.P. Trembley and R.Manohar
Publisher by McGraw hill international editions pp 37-49.
11. Fabio Aioli, Ricardo Cardin, Fabrizio Sebastiani, Alessandro Sperduti,”Preferential Text Classification: Learning Algorithms and evaluation measures”, Springer – Inf Retrieval 2009.
12. Mikolajezyk and C.Schmid “Scale and affine invariant interest point detectors” international journal of computer vision vol 1 2004.

