IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

GENERATING FREQUENT PATTERNS BY IMPROVING THE EFFICIENCY OF APRIORI ALGORITHM

[1]L.Mohan, [2]Dr.T.Venugopal

[1]Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India.

[2]Professor ,CSE department, JNTUH College of Engineering, Jagityal, Telangana, India.

Abstract- Association rule mining is considered as main technique in the data mining area. Association rule mining presents associations among items in transaction databases, relational databases, frequent patterns, correlations and other information of warehouses. Frequent pattern mining is one of the active areas in data mining. It takes part in every data mining task such as prediction, classification, clustering, and association assessment. Finding all frequent patterns is time-consuming process because of its vast number of patterns generated. In this paper, frequent patterns are generated in faster way by improving the efficiency of Apriori algorithm to speed up the process.

Index Terms - Association rules, Apriori algorithm, frequent patterns.

I. INTRODUCTION

Data Mining [1] is the field of finding useful and interesting knowledge, such as patterns, associations and structures, from substantial quantity of data deposited in database, valuable data sources and other useful depositories or warehouses. Therefore, because of the large availability of vast amounts of data in the form of electronic media, and the need for turn it into useful and knowledge information for many applications including market analysis, business management, and support decision making. Data mining has developed into a huge allocation of attentiveness in knowledge handling industry in recent years.

Researchers and business analyst perspective towards data mining as an important step of knowledge retrieval process includes an iterative step such as data integration, data maintenance, data choice, data transformation, extraction of patterns. Data mining constantly explore for innovative, unknown, authorized and theoretically useful patterns in huge data sources. Data mining continuously emphasize on discovering unpredicted and hitherto unidentified correlated data. Data mining uses a task which is a multi-disciplinary one that uses artificial intelligence, machine learning, arithmetical and other data archives. Data Mining can be used for different application areas such as marketing analysis, deceptive finding process, and research conclusions. Data mining process can be termed as procedure of retrieving knowledge discovery in database, information mining, pattern inspection, procuring data. Data mining, the mining of concealed logical data from immense databases, is a widespread method with major capability to help firms focus over utmost information in given data warehouses. Data mining technology or procedures forecast forthcoming trends and behaviors, especially allowing companies to build hands-on as well as knowledge-driven decision making. The mechanized and potential studies delivered by data mining attempt to move further than the studies of past occurrences.

Data mining is called knowledge discovery in databases (KDD) and renowned field for studying database as part of research. The data mining area efficiently finds appealing rules from huge collections of data from the database. The area of Data mining is the procedure of discovering associations or useful patterns among lots of areas in huge relational databases. Association rule is nothing but correlation and the main technique of data mining. The data created by day-to-day activities results in increasing the volume of the data in a database. Therefore, discovering association rules from gigantic amount of data in the database is vital for many firm's decision-making processes includes Market basket analysis and cross marketing. The resultant data helps to discovery the association correlation between the huge number of items generated in database and its typical application to discovery latest valuable rules from the available sales transaction database to find the customer buying behavior patterns just like the impression on the other possessions when buying a specified kind of products or goods. The discovered rules must be used in various areas includes storage capacity planning, sales, customer spending analysis and categorizing the users according to the

purchasing patterns. The identified techniques for finding correlation from the data have conventionally aimed on discovering associations among items stating certain aspects of customer behavior and also purchasing behavior for ascertaining items that buyers purchase jointly. All rules of this kind designate a particular internal pattern and can be effortlessly clarified and expressed through a set or group of association rules.

The amount of obtained information and knowledge can be a source for various applications extending from production control, business/firms management, and market analysis includes customer purchasing behavior from a supermarket or a stock market to engineering automation process and science exploration [2]. As per the downward closure lemma, the items which are generated are identified as candidate set which includes the entire frequent k-length items and next it starts to scan the required transaction database to establish the frequent items between the candidate's items. Therefore, the search for frequent patterns found in every domain. The ideal application is termed as market basket analysis to discover the items most frequently purchased together at general grocery shop or supermarket by studying the customer shopping behavior instantly by using the carts of the customer. Then, after discovering the frequent items they facilitate to obtain association rules between the items and determine about how two items are likely to co-occur in the process of discovering the frequent items.

II. LITERATURE REVIEW

Association rule mining procedure is an association among two or more properties. The goal is to mine association rule to extract every possible rule that are above based on the user specified minimum support and confidence threshold value. Association rule mining [3] discovers stimulating associations and connections between large sets of data items from the given data sets. This association rule technique delivers about how an itemset frequently appears in a specified transaction in a classic example called market basket analysis. The ideal market basket analysis is considered as the essential procedures utilized by huge relationships of the transaction to show associations among numerous items. Market analysis delivers vendors the elasticity to recognize relations among the products that customers buy together frequently in a super market. In a given set of transactions, to uncover rules that will absolutely calculate the existence of an items based on the existences of additional items in the transaction. The Association rule is most useful in analyzing datasets from the given transactions. More often, the data is gathered by means of scanning the barcodes in supermarkets. Therefore, these databases contain a huge number of transactional files that contains all the items which are purchased in single order by the customer. The supermarket manager might recognize and evaluate the specific collections of items are steadily bought together or not. The manger's knowledge is utilized for modifying store arrangements and cross-selling the items and advertisings are based on statistical data.

The purpose of deriving association rules from data sets of a transaction firstly discussed in [4] and termed as market-basket analysis problem. The objective is to give a set of data items and a huge collection of transactions which are said to be the sets (baskets) of items. The aim is to uncover relations between the containments of many items of the transaction data sets within the baskets. The main task in mining of association rules contains discovering all rules that needs to fulfill the user specified limitations on minimum support and confidence with respect to a specified dataset of the transaction. Most routinely utilized association rule mining algorithm used for frequent items sets approach is Apriori algorithm [5].

Apriori algorithm is proposed in 1994 by R. Agrawal and R. Srikant for discovering frequent itemsets in a transaction dataset for extracting association rule. The name itself states that it uses prior knowledge of frequent data itemset properties of a given transactional database. Therefore, need to utilize a circular method or level-wise quest, where k-frequent itemsets are about to find the k+1 itemsets. To increase the efficacy in discovering association rule for the generation of level-wise frequent itemsets, for this an essential property known as Apriori property is used, this property of Apriori reduces the search interval in the data sets. The property of Apriori is that all non-empty subset of frequent itemset should be frequent. Apriori algorithm states that all generated subsets of a frequent itemset should be frequent. If in case the itemset which is generated is infrequent, then all its supersets are going to be infrequent. In the process of discovering frequent items sets, Apriori is the first algorithm designed for association-rule mining. Apriori is considered as the improvement over the other algorithms like SETM and AIS algorithms [6]. Apriori algorithm will search for huge items sets; initially database processed and then utilizes its outcome as the input for finding huge datasets throughout later processes. The results obtained by support level greater than the minimum threshold value is known as huge or frequent items sets and those lower than are termed small items sets. The Apriori process is truly a work based on the huge items set property that define at any subset of a large items set is large, and if a given itemset is not big enough then any of its supersets are not big.

Analysing the entire data that is collected from the database is definitely required for a firm because by using the all data a firm can take proper decisions in extracting useful information. For a firm it is more important to provide information to the users which is useful in assessing the data for analysis and to improve decision making more effectively. One more aspect is to extract and find the entire significant hidden patterns that occur frequently in a data set. Therefore, this paper analyses a frequent pattern mining algorithm called Apriori algorithm to provide an overview of the frequent pattern mining.

III. ASSOCIATION RULE

Association rule mining technique explores interesting patterns and associations between items from the data set of the transaction database. Let us consider an archive of selling transactions, where it is needed to find essential associations between items in a way such that the existence of certain items of a given transaction database will infer the existence of several other additional items of a transaction. The ideal aapplications which includes market basket data analysis, cross- selling in marketing, catalogue design, web log analysis using web mining technique, fraud detection in a system and many more. An association rule example is shown in below figure 1.

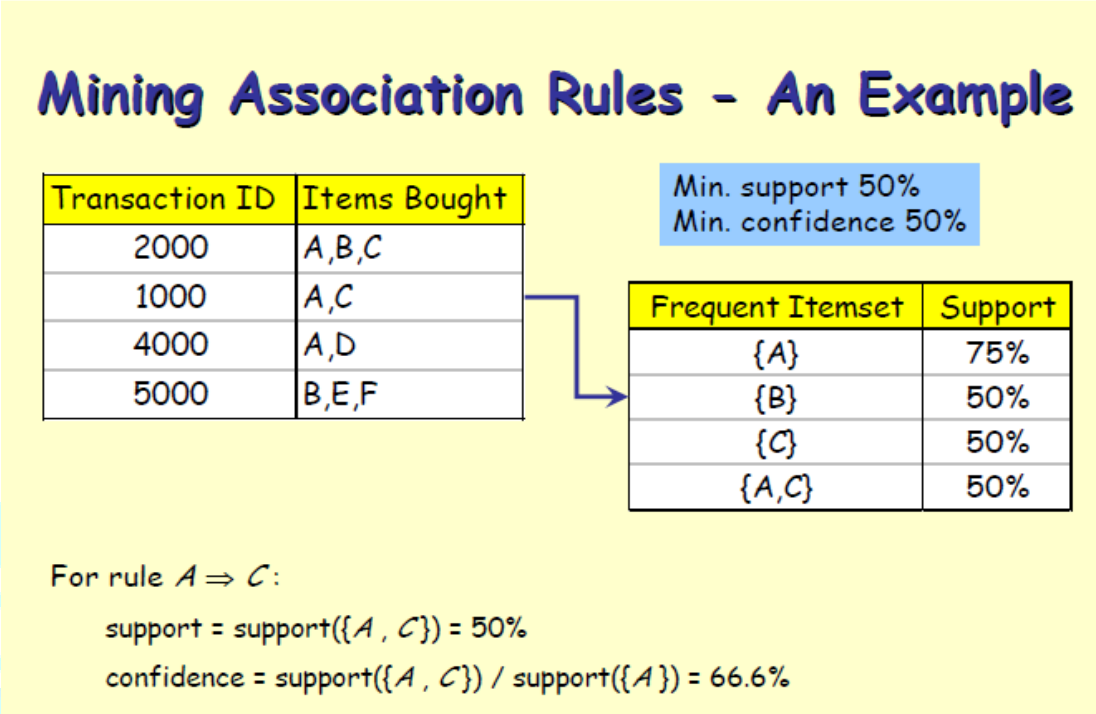

Figure1. Mining Association rule example.

IV. APRIORI ALGORITHM: MINING FREQUENT ITEMS

Apriori Algorithm Working process:

1. Scan the items for support of length 1, called 1- itemsets by searching the database. Remove items that are under minimum support count value.
2. The procedure is to expand the large 1-itemsets into 2-itemsets by joining one item every time to produce all candidate itemsets of length two from the given data sets of the transaction. During this once more scan the support of all candidate itemsets.
3. Reiterate steps 1 and 2 at step k, the formerly discovered (k-1) itemsets expanding into k-itemsets are searched for minimum support count value.

A. Apriori Algorithm

Join Step: Ck is generated by joining Lk-1 with itself.
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.
Ck: Candidate itemset of size k
Lk: frequent itemset of size k
Step 1: L1 = {frequent items};
Step 2**: for** (k = 1; Lk! =∅; k++) **do begin**
Step 3**:** Ck+1 = candidates generated from Lk;
Step 4**: for each** transaction t in database do
Step 5**:** increment the count of all candidates in Ck+1 that are contained in t
Step 6**:** Lk+1 = candidates in Ck+1 with min_support
Step 7**: end**
Step 8**: return** L = ∪k Lk;

Apriori algorithm is the most classical and main algorithm for mining frequent items, proposed by R.Agrawal and R.Srikant in 1994 [6]. Apriori property is used to discover the entire frequent items in a given database D [7]. The main process of Apriori algorithm is to formulate several passes over the given database D. Apriori algorithm applies a circular approach called breadth-first search through the provided search space to explore (k+1)-items using k-items. The function of Apriori algorithm purely determined by the Apriori property called "All nonempty subsets of a frequent itemset must be frequent". There is no doubt that Apriori algorithm effectively [8] discovers the frequent items from the database D as shown in below figure 2. But as the dimensionality of the database [9] increase with the number of items then:

- Extra search space is needed and input/output cost will increase.
- As number of scans in database increased, candidate generation will also increase the results and escalate the computational price.

Frequent Pattern Mining challenges:

- Vast candidate generations
- Scanning transaction database multiple times
- Support counting for candidates having Uninteresting workload.

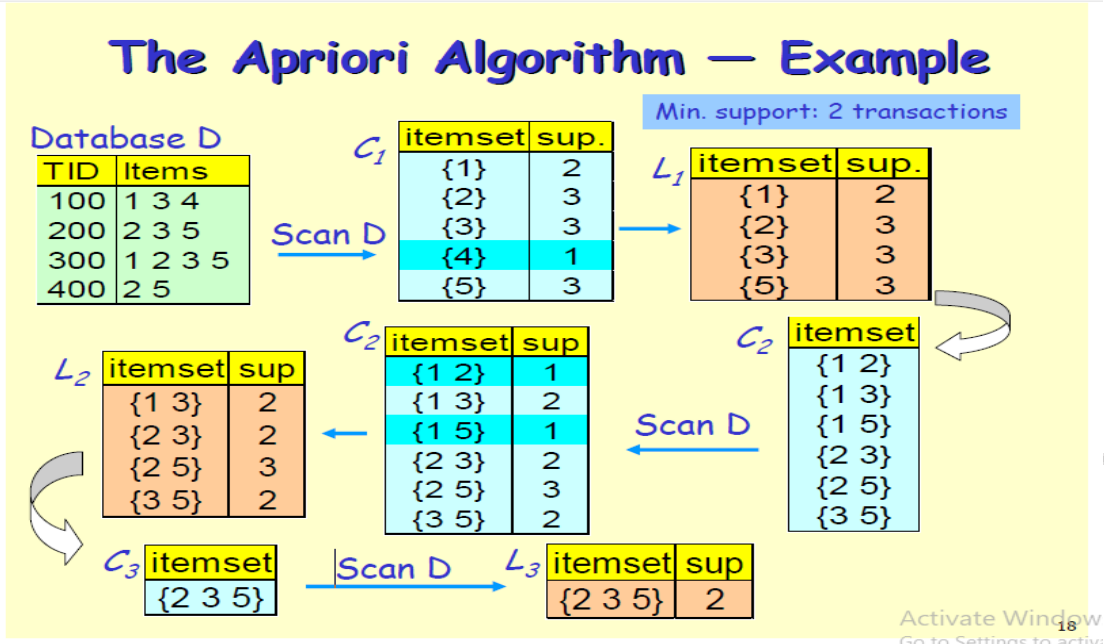

Figure2. Association rule using Apriori algorithm for transaction database D.

B. Improving Apriori Efficiency

The important issue with Apriori algorithm i.e., each pass goes over the given database D.

1. Apriori TID: generating candidates set, in first pass the database is used for counting support only. After each pass it generates more items from the data base for that it needs required additional memory than Apriori algorithm and builds a storage set C^k as shown in below figure 3. The additional memory which is available to store the frequent itemsets per transaction generated.

2. Apriori Hybrid**:** Apriori is used in initial passes as the process initiate and estimate the volume of generated C^k. when C^k is expected to be fit in memory move to Apriori TID.

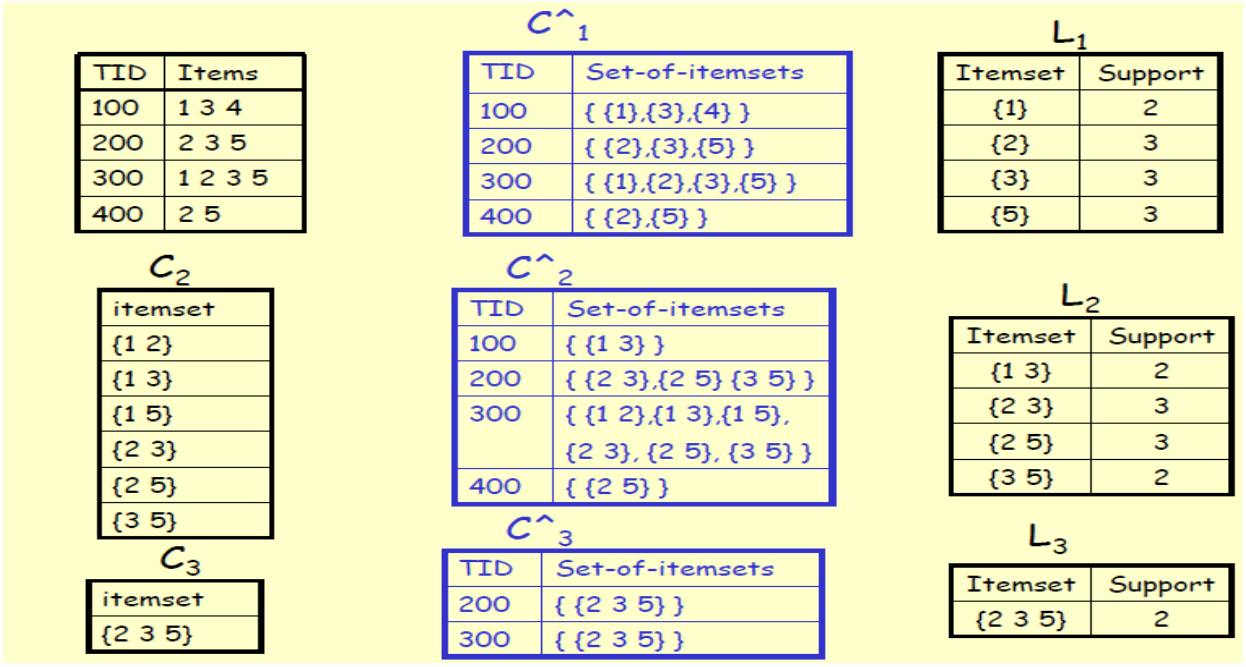

Figure3. Storage set C^k.

1. Transaction diminution**:** The transactions which are generated in previous pass is unusable in subsequent scans, if doesn't have any frequent k-itemset.
   - Reducing the scanning of transaction database.
   - Candidates count should be minimized.
   - Candidates support count should be enabled.

2. Sampling: extracting on a subset of given data.
   - To minimize the probability of loosing some itemsets use the minimum support threshold value.
   - To define the actual itemset count using the rest of the database.

3. Dynamic Itemset Counting (DIC):

The database is partitioned into some blocks, and each block is indicated by a start point.

- Dynamic Itemset Counting assesses the support of entire itemsets that are presently considered and joins newly generated itemsets to the set of candidates itemsets, if entire subsets are assessed to be frequent at each start point.

- If in case Dynamic Itemset Counting joins all frequent itemsets to the set of candidates itemsets for the duration of the first scan and have counted every itemsets precise support at certain point for the duration of the next scan. Thus, Dynamic Itemset Counting can finish the process of scanning in two scans.

V. CONCLUSION

Association rule possibly the very important contribution in database community in a Knowledge Discovery Database. Many interesting topics have been researched. Association rule is used to find correlations, associations, frequent patterns, or causal structures between sets of items in transaction databases. To understand consumer purchasing behaviour by discovering correlations and associations among the distinct cart items that buyers place in shopping cart. This paper, elaborate the work presented on association rule mining technique by improving the efficiency of Apriori algorithm to efficiently discovery frequent patterns from the large databases in a faster way.

REFERENCES

[1] Tan, P. N., M. Steinbach, V. Kumar, "Introduction to data mining", Addison-Wesley, 2005, 769pp.
[2] H. Grosskreutz, B. Lemmen, and S. Reuping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, vol. 4, no. 3, pp. 147-165, 2011.
[3] Agrawal, R., Imielinski, T. and Swami, A. "Mining association rules between sets of items in large databases", In Proceedings of the International ACM SIGMOD Conference, pages 207–216,1993.
[4] Kotsiantis S, Kanellopoulos D., "Association rules mining: a recent overview", GESTS International Transactions on Computer Science and Engineering, Vol.32 (1),2006, pp. 71-82.
[5] R. Jin and G. Agrawal, "An efficient implementation of Apriori association mining on cluster of SMPs", Proc. Workshop High Performance Data Mining (IPDPS 2001), Apr. 2001.
[6] Agrawal R, Srikant R., "Fast algorithms for mining association rules", VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

[7] Rakesh Agrawal and Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc 20th International Conference Very Large Data Bases (VLDB), pp. 487-499, Year 1994.
[8] Mannila H, Toivonen H, Verkamo A I., "Efficient algorithms for discovering association rules" AAAI Workshop on Knowledge Discovery in Databases (SIGKDD). July 1994, Seattle, 181-92.
[9] Secure Mining of Association Rules in Horizontally Distributed Databases, TamirTassa , IEEE Transactions On Knowledge And Data Engineering, VOL. 26, NO. 4, April 2014.