Mining Recurrent Pattern Identification on Large Database

Shivangi Srivastava
M.Tech, Computer Science
Amity University
Noida, India
shivangi100@gmail.com

Ganesh Khadanga
Technical Director
National Informatics Center
New Delhi, India
ganesh@nic.in

Divya Gupta
Assistant Professor
Amity University
Noida, India
fromdivya81@gmail.com

***Abstract*—Recurrent pattern mining is an important problem in the context of data mining. In this paper data mining algorithms have been discussed and compared. Recurrent pattern mining has been an important area in data mining research and it is the first step in the analysis of data rising in a broad range of applications. The algorithms are compared with respect to the items like methodology and its basic principles in terms of the elements user like support, and scan of the database (full or partial).**

Keywords: Frequent item sets, Apriori, Association rule, Support, CDS

I. INTRODUCTION

A recurrent pattern is a set of items, subsequences, substructures, etc. which occurs frequently in a data set. It is the most powerful problem in association mining. Data mining, or the efficient discovery of interesting patterns from large collections of data. Association rule mining is a significant data mining technique to generate correlation and association rule. An association rule is of the form A=> B, where A c I, B c I and A ∩ B =ϕ. The rule A =>B holds in the transaction set D, with support supp., where supp. is the percentage of transactions in D that contain A U B (i.e, the union of sets A and B, or say , both A and B). This rule is taken to be the probability, P(AUB). The rule A =>B has confidence c in the transaction set D, where c is the transaction percentage in D containing A that also contain B.

Support (A=> B) = P(A U B)

Confidence (A=> B) = P(B|A)= Support(A U B)/Support(A)

Rules should satisfy minimum support and minimum confidence in order to determine recurrent item sets. Recurrent pattern mining is a two step process:

Find all frequent item sets i.e. each of these item sets will occur at least as frequently as a predetermined minimum support count.

Initiate strong association rule from frequent item sets i.e. these rules must satisfy minimum support and minimum confidence.

In this paper data mining algorithms have been discussed and compared and best algorithm has been chosen.

Data Mining Algorithms

Number of algorithms are available for data mining. In this paper we have taken up the Apriori Algorithm, Compacting Data Set (CDS), Frequent Pattern Algorithm using Dynamic Function, Multilevel association rule mining algorithm based on Boolean matrix and the Frequent Pattern Growth Algorithm for the study and comparison. All the above algorithms were examined with respect to their basic principle and suitability.

1. Apriori Algorithm –It is a seminal algorithm for mining frequent item sets. This algorithm uses the prior knowledge frequent item set properties and a level wise search. The algorithm prunes many sets which are unlikely to be the frequent set before reading the database. In the first pass the algorithm counts the item occurrences to find the frequent items. Subsequently the joining and pruning process is carried out. In the joining step the candidate k item set is generated by joining k-1 item itself. In the pruning step a database scan to

determine the count of each candidate set that satisfy a count number less than the minimum support count are finalized.

The key of the apriori mining association rules is to fix the appropriate support and confidence values to find frequent itemset. All the other algorithms have introduced new concepts as an improvements over the apriori and attempted to bring efficiency and reduced database scan.

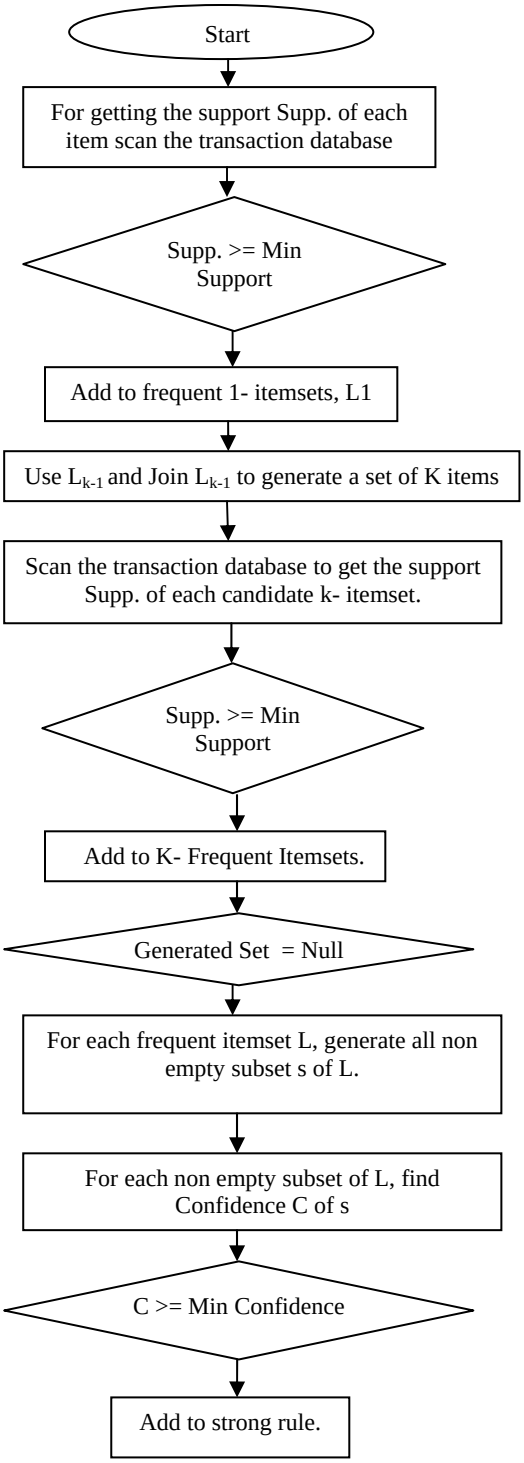

Fig 1: Flow Chart Apriori

2. Compacting Data Sets – In this approach first duplicate transactions are being merged and then intersection between item sets is done and deleting unneeded subsets repeatedly[1]. This algorithm is different from all classical frequent itemset discovering algorithms in such a way that it not only removes unnecessary candidate generation but also removes duplicate transactions.

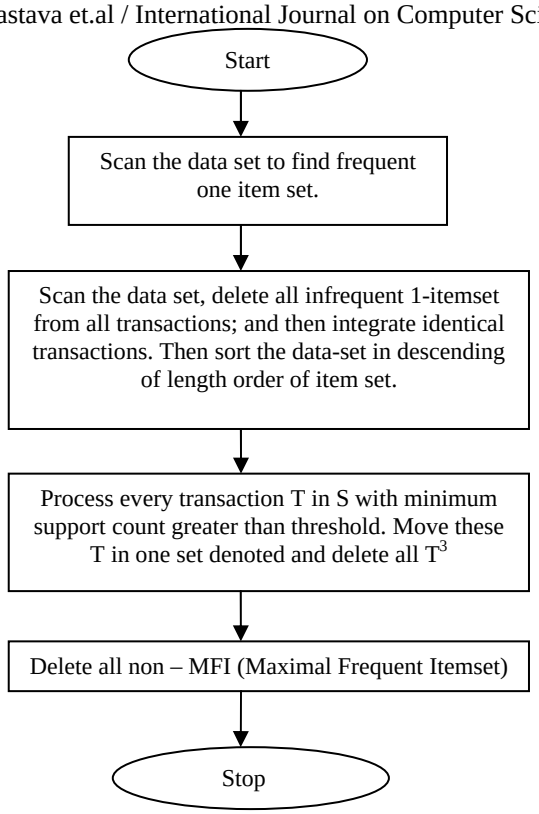

Fig 2: Flow Chart Compacting Data Set

3. Frequent Pattern Algorithm using Dynamic Function - This algorithm scans through the entire database and transaction pairs are generated with longest common sequences and computes longest common sequences of item id for each previous transaction pair. Then the algorithm prunes the transaction pairs with empty longest common sequences. The longest common sequence is found using the dynamic function. The support count is done for pruned subset patterns rather than the whole database. In the next operation again the pruned transaction pair with the least common sequences were observed. The advantage with this approach is that the database access is reduced and the subsequent iteration is faster than the previous iteration.

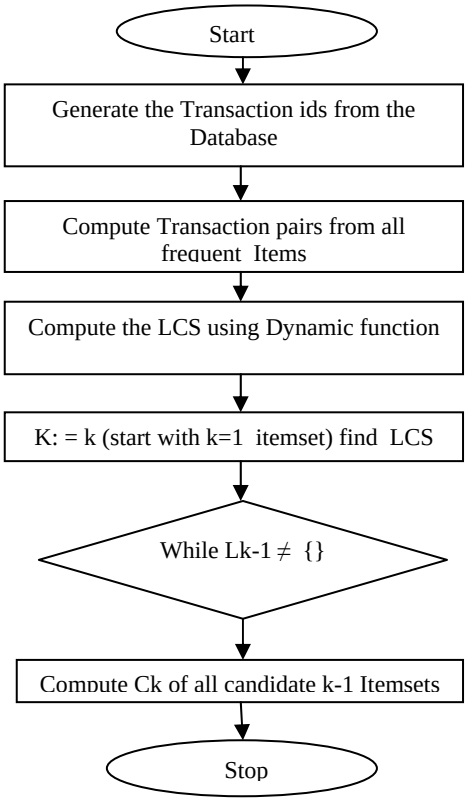

Fig 3: Frequent Pattern Algorithm using Dynamic Function

4. Multilevel association rule mining algorithm based on Boolean matrix -

In this algorithm a Boolean matrix based approach is used to find out the frequent item sets. The algorithm scans the database once and prepares the association rules. Then the apriori property is used to prune the item sets. The algorithm generates the Boolean matrix in the form of bits for the transactions. The Boolean matrix consists of "0" and "1" and the "and" operation is defined as 0.0=0, 1.0=0, 1.0=0 and 1.1=1. Then the matrix dimension is reduced based on item with minimum support. Then the sum of the element values over the matrix for all two items set. Then the AND operation is carried to generate the 3 items set. Once the maximum frequent itemset is found the algorithm stops. The advantage with this algorithm is that it scans database only once and it needs less memory for the operations.

5. Frequent Pattern Growth Algorithm - In this algorithm a FP growth tree table is prepared from the transaction database using all the transactions order in a descending order after removing the infrequent items from the database. It stores the actual transactions from the database and every item has a linked list. This new structure is identified as a FP tree. This consist of the root node and a set of child nodes and a frequent item header table. Subsequently the node link structure and the insert-tree(P,N) subroutine is used to find out the frequent pattern.

II. COMPARISON OF ALGORITHMS:

The apriori algorithm works only for static database. They have used candidate itemsets generation method, but this approach was highly time consuming [1]. In Compacting Data Sets (CDS) approach first duplicate transaction is merged and then intersection between itemsets is taken and then unneeded subsets are deleted repeatedly. This classical algorithm differs from other algorithms in such a way that it not only removes unnecessary candidate generation but also remove duplicate transactions. The main features of multilevel association rule mining algorithm based on Boolean matrix are that it scans the transaction database once, it does not produce itemsets, and it make use of the Boolean vector "relational calculus" to discover frequent itemset[3]. It stores all transaction data in bits, so it requires less memory space and can be used for mining large transaction databases. In Frequent Pattern Algorithm using Dynamic Function, mining the transposed database runs through a smaller search space. FP growth algorithm mines frequent item sets from FP-Tree without generating candidate frequent item sets unlike Apriori. The major issue of Apriori based algorithm that is the cost to generate candidate frequent item sets has been addressed in FP growth algorithm.

In paper[11] the testing results of experiments have been shown in the figure. In that Figure, the horizontal axis represents the number of support in database and the vertical axis represents mining time. The three curves denote different time cost of the algorithm Apriori, FP Growth and FPMDF(Frequent Pattern Mining using Dynamic Function) with different min support.

Table 1: Comparison of various algorithms

Name of the Algorithm	Features/Principles	Database Scan	Support	And/OR
Apriori	Uses Prior Knowledge of frequent itemset properties K itemsets are used to explore (k+1) itemsets	Large Database Scan	Yes	NA
Compacting Data Sets	Merging of duplicate transactions & intersection between itemsets is taken	Atleast once but less than in apriori	Yes	NA
Frequent Pattern Algorithm using Dynamic Function	Transpose database then result is very fast	Atleast once	No	NA
Multilevel Asso-rule mining algorithm based on Boolean matrix	Uses Boolean logical operation to generate the multilevel association rules & top-down approach	Only once	Yes	Yes
Frequent Pattern Growth	Mines frequent item sets from FP-Tree without generating candidate frequent item sets	Only two	Yes	NA

III. CONCLUSION:

There are number of algorithms for data mining and active research is going on in this field. Each technique has its own pros and cons. Performance of particular technique depend upon input data and available resources[12]. Mining recurrent pattern is efficient method for discovering frequent pattern. It is a well known that the way candidates are defined has great impact on running time and memory need and this is the reason for the large number of algorithms.

IV. REFERENCES:

[1] Nidhi Sethi and Pradeep Sharma. Mining Frequent Pattern from Large Dynamic Database using Compacting Data Sets.International Journal of Scientific Research in Computer and Engineering, May-June 2013, pages 31-34.

[2] Sunil Joshi, Dr. R.S. Jadon and Dr. R.C.Jain. An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function. International Journal of Computer Application, November 2010, pages 37-41.

[3] Pratima Gautam and K.R. Pardasani. A Fast Algorithm for Mining Multilevel Association Rule Based on Boolean Matrix. International Journal on Computer Science and Engineering, November 03, 2010, pages 746-752.

[4] Laboratory Module 8, Mining Frequent Itemsets – Apriori Algorithm.

[5] Frequent Item Set Mining Methods Jiawei Han und Micheline Kamber, Data Mining – Concepts and Techniques, Chapter 5.2

[6] Implementation of the Apriori algorithm in C#.; Author: Omar Gameel, http://www.codeproject.com/Articles/70371/Apriori-Algorithm.

[7] Sandhya Rani Jetti, Sujatha D. Mining Frequent Item Sets from incremental database: A single pass approach International Journal of Scientific & Engineering Research, Vol. 1, No. 2, July-December 2010,pp.433-441.

[8] Pramod S. and O.P. Vyas. Survey on Frequent Itemset Mining Algorithms. International Journal of Computer Applications, pages 94-100.

[9] Goswami D.N, Chaturvedi Anshu. Raghuvanshi C.S An Algorithm for Frequent Pattern Mining Based On Apriori Goswami D.N. et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947.

[10] Sunil Joshi, R S Jadon and R C Jain. A framework for Frequent Pattern Mining Using Dynamic Function, International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011

[11] Vikas Kumar, Sangita Satapathy. A Review on Algorithms for Mining Frequent Itemset Over Data Stream International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, April 2013, pages 917-919.

[12] Deepak Garg, Hemant Sharma. Comparitive Analysis of Various Approaches Used in Frequent Pattern Mining International Journal of Advanced Computer Science and Applications, pages 41-47.

[13] Jiawei Han, Hong Cheng, Dong Xin, Xiefeng Yan, Frequent Pattern Mining: current status and future directions, Springer Science+Business Media, LLC 2007.

[14] Mahmood Deypir, M Sadreddini, & S Hashemi. " Towrds a variable size sliding window model for frequent itemset mining over data streams" Elsevier (2012).

[15] Han Jiawei, Pei Jian, and Yin Yiwen. Mining Frequent Patterns without Candidate Generation. SIGMOD, 1-12, Dallas, TX, May 2000.