© 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
National Conference on Advanced Research Trends in Information and Computing Technologies (NCARTICT-2018), Department of IT, L. D. College of Engineering, Ahmedabad, Gujarat, India In association with
International Journal of Scientific Research in Science, Engineering and Technology

A Survey on Efficient Frequent Pattern Mining Techniques

Janak Thakkar[*1], **Dr. Mehul Parikh**[2]
[*1]IT Department, R C Technical Institute, Ahmedabad, Gujarat, India
[2]Associate Professor, IT Department, LDCE, Ahmedabad, Gujarat, India

ABSTRACT

Data Mining is the technique to abstract the useful data from the large dataset for different perspectives. Frequent pattern mining has become an important data mining technique to find the frequent patterns from the data set that appears frequently. Frequent Pattern Technique is widely used in financial, retail, telecommunication and many more. The major concern of these industries is faster processing of a very large amount of data. Various techniques and algorithms have been proposed for this purpose. Apriori, FP-tree are the pioneer techniques among them. In this paper, we have analysed algorithms for finding frequent patterns with the purpose of discovering how these algorithms can be used to obtain frequent patterns over large transactional databases with most efficient way in various aspects. This has been presented in the form of a comparative study of the following algorithms: Apriori, Frequent Pattern (FP) Growth, dNC-ECPM Algorithm, OCFP–growth, IA-TJ-FGTT(Important Attributes - Transaction Joining - Frequency Gathering Table Technique).

Keywords: Data Mining, Frequent Pattern Mining, Apriori, FP-Growth, Erasable Patterns, Close Patterns

I. INTRODUCTION

Frequent patterns are the itemsets that occur with frequency which is greater than a user-specified threshold. Frequent patterns plays an significant role in mining associations, correlations, and many other interesting relationships among data in datasets. It also helps in data indexing, classification, clustering, and other data mining tasks as well. Thus, FPM is an important data mining technique and topic for the research.

FPM was first introduced by Agrawal etal.[1,2] for market basket data analysis in the form of association rule mining, which is useful for discovering interesting relationships hidden in large data sets. For example, we may find a strong relationship, which can be represented in the form of association rules or sets of frequent items, exists between the sale of diapers and beer because many customers who buy Bread also buy Butter.

A. Apriori algorithm

Agrawal and Srikant [2] noticed an interesting downward closure property, called Apriori, among frequent k-itemsets: A k-itemset is frequent if all of its sub-itemsets are frequent. For this, first scanning the database to give the frequent 1-itemsets, then using the frequent 1-itemsets to generate candidate frequent 2-itemsets, by this process frequent k-itemsets can be generated for some k. This is the essence of the Apriori algorithm[7].Though in many cases, the Apriori algorithm reduces the size of candidate sets. However,
it has two significant disadvantages : (1) generating a huge number of candidate sets, and (2) Scanning of dataset more number of times.

B. FP-growth algorithm

Han et al. [3] Proposed an FP-growth method that mines the complete set of frequent itemsets without candidate generation in less number of Scans. FP-growth follows divide-and-conquer method[8]. In first scan of the database, it derives a list of frequent items in which items are ordered by frequency-descending order. According to the frequency- ascending list, the database is compressed into a frequent-pattern tree (FP-tree), which retains the itemset association information. The FP-tree is mined by starting from each frequent length-1 pattern (as an initial suffix pattern), which constructs the conditional pattern base, then constructing its conditional FP-tree, and performing mining recursively on such a tree. The pattern growth is achieved by the joining of the suffix pattern with the frequent patterns generated from a conditional FP-tree. The FP-growth algorithm transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then joining the suffix. It uses the least frequent items as a suffix, offering good selectivity. Performance studies shows that FP-Growth method significantly reduces search time.

The remaining document is organised as follow: In Section 2 will represent three different algorithms for the mining of frequent patterns in more efficient way. In section 3, Comparative study of all this algorithms will be represented in tabular format which covers some basic characteristics of algorithms with significant advantages and disadvantages. Section 4 will cover the conclusion with some limitations and future work. In section 5, all references will be listed.

II. LITURATURE SURVEY

As Apriori and FP - Tree have some disadvantages, to overcome this paper proposed some novel algorithms and techniques for efficient pattern mining.

C. dNC-ECPM Algorithm

Tuong Le, Giang Nguyen, Tzung-Pei Hong [4] proposed an efficient algorithm for erasable closed pattern from dataset to reduce Memory Usage and Mining Time. Paper mainly focus on the Erasable Closed Patterns (ESP) from the dataset. In algorithm first step is to identify the ECPs from the dataset in such a way that which doesn't affect the information loss. In second step dNC-ECPM algorithm will be implemented for mining the ECPs generated in first step. The dNC-ECPM algorithm follow the systemic approach to discard the erasable close patterns. Initially dNC_set structure is defined which consists set of diff. Node Codes Based on WPPS–tree(Weight, Pre order, Post Order, Childnodes). After successful generation of dNC_set structure, algorithm is implemented. The NC_diff algorithm is to determine the dNC_set and the relationship of dNCs(X) and dNCs(Y). A node code Ci is an ancestor of another node code Cj if and only if Ci.pre-order ≤ Cj.pre-order and Ci.post-order ≥ Cj.post-order. The dNC_set algorithm to construct WPPC-tree with threshold value and determine E1 and generate dNCset of erasable patterns from E1. Significant advantage of the dNC_ECPM is less memory usage and mining time.

D. OCFP – Growth Algorithm

Another approach for efficient frequent pattern mining was proposed by team of Hsiao-Wei Hu, Hao Chen Chang, Wen-Shiu Lin [5]. To overcome the limitations of Apriori algorithm OCFP–growth (Optimized Close Frequent Pattern) algorithm was proposed. The algorithm is based on the OMIS – tree (Optimized Minimum Item Support Trees). In this algorithm for Minimum support on the individual items will be counted and after that minimum support on the itemset. Definition 2: A subset of frequent itemset may be not frequent. Definition 3: A subset containing the item with the lowest MIS value in a frequent itemset must be frequent. The process of to give input OMIS-tree, which is a set of frequent item F,MIS of each item in F, where the value of k = 2. Output will be the complete set of all f's conditional frequent k-itemset (patterns). After input is taken the step by step functions will be implemented. In first step OCFP-growth algorithm will be called with MIS-tree by declaring root as a null. Second step is to generating the patterns with MIS value. In third step, set of the conditional patterns will be constructed. Fourth and final step is to add frequent patterns will be added into the MIS Tree and run the programs till all infrequent patterns are discarded.

E. IA-TJ-FGTT Technique

Another technique is developed by the Saravanan. Suba, Dr.T. Christopher[6] to mine the frequent patterns from the large datasets. Algorithm is IA-TJ-FGTT which stands for Important Attributes -Transaction Joining-Frequency Gathering Table Technique. The technique is based on the association rule[7][8] mining method which perform pruning, joining operations on the attributes. In techniques first important attributes are selected from the all attributes. After selecting the important attributes duplicate transactions are discarded and dataset is reduced. Then transactions are joined. After joining the truncation Frequency Gathering Table(FGT) is generated which resulting in the Frequent Patterns.

III. COMPARITIVE STUDY

After evaluating individually now below listed table shows the comparative analysis of the all the approaches discussed in the paper. comparison is focused on the selected fields like, Technique used by the algorithm, Advantages, Disadvantages, Storage structure, Mining Time and Accuracy.

Table 1. Comparative Analysis of various Pattern Mining Algorithms

	Apriori	**FP -Growth**	**dNC-ECPM**	**OCFP**	**IA-TJ-FGTT**
Technique	Breadth First Search	Divide And Conquer	Divide And Conquer	Closed Frequent Pattern	Association Rule Mining
Advantages	- Easy to Implement	- Scanning of Dataset for Two times only	- Memory Usage is low	Execution time is less	- get perfect patterns
Disadvantages	- Too many scans of dataset - Require Large Memory Space	- expensive then the Apriori	- not efficient in case of Large Datasets	- generating the MIS tree is difficult	- mining time is more compare to others because of more operations
Storage Structure	Array	Tree	Tree	Tree	Array
Mining Time	More	Less Compare to Apriori	Less than Apriori and FP -Growth	Reduced Mining Time and Execution	More as compare to dNC-ECPM and OCFP
Accuracy	Less Compare to others	Less Compare to others	High in Small Dataset Low in Large Data Set	Better than other	High than all the Techniques

IV. CONCLUSION

In this paper, we gave the brief overview of some of the existing algorithms and new approaches for the frequent pattern mining. Than each individual algorithm is compared and analyzed with the remaining algorithms by various fields. The comparison shows that all the approaches are efficient in reducing mining time but up to some extent not effective in the accuracy. More focused research can be possible in the direction of getting accurate and concrete frequent patterns which can be utilised for the planning and forecasting.

V. REFRENCES

[1] First Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACMSIGMOD international conference on management of data(SIGMOD’93), pages207-216.1993

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules. InProceedings of the 1994 international conference on very large databases(VLDB’94), pages487–499.1994

[3] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In Proceeding of the 2000 ACM-SIGMOD international conference on management of data (SIGMOD’00) pages1–12.2000

[4] Efficient algorithms for mining erasable closed patterns from product datasets” 10.1109/ACCESS.2017.2676803, IEEE Access

[5] “An optimized frequent pattern mining algorithm with multiple minimum supports” 2016 IEEE International Conference on Big Data (Big Data) Hsiao-Wei Hu, Hao-Chen Chang, Wen-Shiu Lin

[6] Saravanan. Suba, Dr.T. Christopher. "An Improved and Efficient Frequent Pattern Mining Approach to Discover Frequent Patterns among Important Attributes in Large Data set Using IA-TJFGTT” 2016 IEEE International Conference on Advances in Computer Applications (ICACA)

[7] Goswami.D, Chaturvedi.A, Raghuvanshi.C “An Algorithm for Frequent Pattern Mining Based On Apriori” IJCSE 2010.

[8] Mabroukeh.N and Ezeife.C “A Taxonomy of Sequential Pattern Mining Algorithms” ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date, November 2010