RESEARCH PAPER | Engineering | Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X



A Study and Analysis of Techniques to Mine Frequent Itemsets

KEYWORDS: Frequent, K-itemset, Mine

Patel Atul R.	Prof. Tushar S. Patel
ME (CSE Student), S. P. B. Patel Engineering College, Mehsana, Gujarat, India	Asst. Professor, IT Department, S. P. B. Patel Engineering College, Mehsana, Gujarat, India

ABSTRACT Data mining refers to extracting knowledge from large amounts of data. Frequent itemsets is one of the emerging task in data mining. An Apriori algorithm to two limitation first generate huge candidate itemsets and second more times scan the database. Problem to be solved some method for frequent itemset mining in paper. Three major factors used in frequent itemset mining such as time, scalability, efficiently. In this paper analyze the various methods for frequent itemset mining such like TDTR, DAPS, MBAT, TABULAR, DIC and FREQUENT K-ITEMSET also discuss advantages & disadvantages of the frequent itemset mining algorithm .

I. INTRODUCTION

Data mining is the process of discovering interesting knowledge from large amounts of data stored in database, data warehouse, or other information repositories. Popular research area of data mining was started by the tasks of frequent item set mining and association rule induction. The huge research efforts devoted to these tasks have led to a variety of sophisticated and efficient algorithms to find frequent item sets. Among the best-known approaches are Apriori, Eclat and FP-growth [7]. Frequent pattern mining is the process of searching recurring relationships in a given dataset. Frequent patterns are patterns (i.e. itemsets) that appears in a dataset frequently. A set of items, i.e. computer and antivirus that appears frequently together in a transaction dataset is a frequent itemsets. Frequent patterns mining like Frequent itemsets find frequent itemsets from the small database and/or large database, where the database are either transactional or relational. The frequent itemset mining is the process of finding out frequent itemsets from the DB. Frequent itemsets such like 1-frequent, 2-frequent, 3-frequent. . . . . k-frequent itemsets.

Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems. Frequent item set mining is a data analysis method, which was originally developed for market basket analysis and which aims at finding regularities in the shopping behavior of the customers of supermarkets, mail-order companies and online shops [7].

Frequent pattern mining was first proposed by Agrawal et al. (1993) for market basket analysis in the form of association rule mining. It analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets". For instance, if customers are buying computer, how likely are they going to also buy antivirus (and what kind of antivirus) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and arrange their shelf space. Association rules describe how often items are purchased together.

The frequent itemset mining algorithms such as, Apriori, AprioriTID, DHP, FDM, DIC, PincerSearch, CARMA, Eclat, Diffset, FP-growth , DSM-FI, PRICES, FHARM, H-mine, DTFIM, GIT-tree, Scaling Apriori, FPG ARM etc.

In this paper, section II discuss the problem definition of the frequent itemset mining; section III discuss the various methods for frequent itemset mining; section IV discuss analysis and discussion; Finally section V concludes the paper.

II. PROBLEM DEFINATION

Generate the huge number of candidate itemsets and more time scan to database. Also the major challenge found in frequent pattern mining is a large number of result patterns. This problem further to be solved that various methods for frequent itemset mining. The main aim of the frequent itemset mining to optimize the process of finding patterns which should be efficient, scalable and detect the important patterns.

III. METHODS

1) TDTR Algorithm

A Two Dimensional Transaction Reduction technique was proposed to reduce the number of transactions from the original database (called Transaction Reduction) to find the frequent itemsets. This algorithm reduces the number of scans in the database and improves the efficiency and reduces the execution time. Also reduce the number of scan from the DB. Find the largest frequent itemsets using the two dimensional approach. In TDTR technique generate the large frequent itemsets and after generate 1-frequent itemsets, 2-frequent itemsets and so on. The detail parameters are shown in Table I.

2) DAPS Algorithm

A Dynamic Approach Pattern Search technique .In this technique first scan the DB. And to be convert in to vertical format database i.e. (Item Id, TID). Count the number of items in each transaction associated to Count1. And check out the condition if Count1 less then equal to minimum support then remove the transaction in to DB. Generate the 1-frequent itemsets (L1) after find the two candidate itemsets. Make the table for two itemsets and that count the number of item in transaction from the DB. Check the condition if Count2 less then equal to minimum support then remove the transaction in to DB generate the 2-frequnet itemsets (L2). Continue the procedure to find the k-frequent itemsets (Lk) and generate the large frequent items. This

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

technique used to reduce the transaction in to DB. The detail parameters are shown in Table II.

3) DIC Algorithm

A Distributed Itemset Counting technique used to reduce the number of DB. Dynamic Itemset Counting which represents a different approach to frequent itemset generation in a distributed ARM environment such as OPT-DIC (optimistic messaging). It is based upon the downward disclosure property in which adds the candidate itemsets at different point of time during the scan. In this dynamic blocks are formed from the database marked by start points and unlike the previous techniques of Apriori it dynamically changes the sets of candidates during the database scan. Unlike the Apriori it cannot start the next level scan at the end of first level scan, it start the scan by starting label attached to each dynamic partition of candidate sets [8]. The detail parameters are shown in Table III.

4) MBAT Algorithm

A Matrix Based Algorithm with Tags to finds the frequent itemsets directly from the transactional matrix which is generated from the database to generate the association ruls, reduces the number of candidate itemsets, mainly candidate 2-itemsets and reduce the transaction in to DB. This technique includes first transaction DB in to transaction matrix with tag format. Second find out the 2-frequent itemsets, 3-frequent itemsets,.....k-frequent itemsets with transaction reduction formula. Also generate the candidate itemsets. The detail parameters are shown in Table IV.

5) Tabular Algorithm

A Tabular method to generating the frequent itemsets and reduce the number of transaction. It does not generate the candidate itemsets. In this method first scan the DB and create a initial table i.e. Items as rows and Transactions as columns. Item is present indicate "1" otherwise "0" . Second count the number of items presented in each row as RC and count the number of "1" in each column as CC. Generate the frequent itemsets, does not generate the candidate itemsets and reduce the transaction based on minimum support. The detail parameters are shown in Table V.

6)K-itemset Algorithm

A K-itemset method to generating the frequent 1- itemsets and frequent k- itemsets. In this method, 1-itemset will be generated directly from the transaction database. To find the frequent K-itemsets. The algorithm starts the searching of itemsets whose length is at least K, i.e., the itemsets whose length is less than K will be pruned from the transaction database which reduce the size of the database and further processing overhead. It generates 1-itemsets as a data pre-processing step which save time and makes execution fast. The detail parameters are shown in Table VI.

TABLE I
TDTR ALGORITHM PARAMETERS

Storage Structure	Array based
Technique	Transaction Reduction technique apply to reduce the number of transactions from the DB
Memory Utilization	Require less amount of memory
Databases	Suitable for sparse dataset, medium dataset, dense dataset
Time	Execution time is less where minimum support is low

TABLE II
DAPS Algorithm Parameters

Storage Structure	Array based
Technique	Based upon dynamic insertion of candidate items. This technique to improve the efficiency and scalability for Apriori.
Memory Utilization	Require different amount of memory at different point of time
Databases	Suitable for sparse dataset, medium dataset
Time	Execution time is small because dynamic itemset are added according to situation

TABLE IV
MBAT ALGORITHM PARAMETERS

Storage Structure	Array based
Technique	Find the frequent itemsets directly from transaction matrix, reduce the transaction, and reduces the number of candidate itemsets mainly candidate 2-itemsets
Memory Utilization	Require less amount of memory
Databases	Suitable for medium dataset and dense dataset
Time	Execution time is very less compare to Apriori

TABLE V
TABULAR ALGORITHM PARAMETERS

Storage Structure	Array based
Technique	Find the frequent itemsets, reduce the transaction .This technique does not generate the candidate itemset compare to Apriori
Memory Utilization	Require more amount of memory
Databases	Suitable for sparse dataset, medium dataset and dense dataset
Time	Number of transaction is increase then time is increase

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

TABLE VI
K-ITEMSET ALGORITHM PARAMETERS

Storage Structure	Array based
Technique	Find the 1- frequent itemsets and k-frequent itemsets
Memory Utilization	Require much amount of memory
Databases	Suitable for sparse dataset, medium dataset
Time	At minimum support of 2%, k-itemset takes less time than Apriori for finding the frequent 3-itemsets.

VI. ANALYSIS AND DISCUSSION

We have study about various methods such as, TDTR, DAPS, DIC, MBAT, Tabular and Frequent K-itemset for frequent itemset mining. The advantages and disadvantages of various methods are shown in Table VII

TABLE VII
ADVANTAGES AND DISADVANTAGES

Methods	Advantages	Disadvantages
TDTR	(1) Reduces the transactions (2) Reduces the number of scan in database (3) Improves the efficiency and reduces the execution time (4) Find optimal solution	(1) Using only two-dimensional array
DAPS	(1) Find longest common subsequence (2) Less memory use (3) Reduces the number of scan in DB (4) Improve efficiency and reduce the computing time	(1) Must be Convert the Transaction database into Vertical format database
DIC	(1) Improve the efficiency of Apriori (2) Reduces the database scan (3) Reduces the candidate itemsets	(1) Not mining frequent itemsets in non-distributed environment
MBAT	(1) Database scan only once compare to Apriori (2) Reduce large number of candidate itemsets (3) Reduce the transactions (4) Reduce the execution time (5) Frequent itemsets generate directly from transaction matrix	(1) Tags must can be used
TABULAR	(1) Reduce the transactions (2) Reduce number of scan in DB (3) Execution is medium	(1) Generate the number of tables
K-ITEMSET	(1) Provide high accuracy and efficiency (2) Reduces the size of the DB	(1) Only generate the 1-frequent itemsets and k-frequent itemsets (2) Require more number of tables

VI. CONCLUSION

The analyze of some methods for frequent itemsets mining. TDTR method better then the DIC method because TDTR reduce the transcations. When DIC does not reduce the transaction its reduce the candidate itemsets. Also TDTR require less amount memory compare to DIC. Tabular method better then DAPS method because in Tabular method reduce the transaction when DAPS method does not reduce the transaction. Also in DAPS method must be converted from horizontal data format to vertical data format.

REFERENCE [1]D.Kerana Hanirex," A Comparative Analysis on Mining Frequent Itemsets," Proceedings of International Journal of Advanced Research in Computer Science and Electronics Engineering, Volume 1, Issue 10, December 2012. | [2]Birla Sunderlal and Prof.Anju Singh, "An Efficient Method for extracting Frequent Pattern Using Transposition of Database." Proceedings of the Journal of Current Engineering Research,21-26,November-December 2012. | [3] Preeti Paranjape and Dr.Umesh Deshpande," A DIC-based Distributed Algorithm for Frequent Itemset Generation." Journal of Software, Vol. 6, February 2011.
[4] Harpreet Singh and Renu Dhir," A New Efficient Matrix Based Frequent Itemset Mining Algorithm with Tags," Proceedings of the International Journal of Future Computer and Communication, Vol. 2, August 2013. | [5] G.Krishna Mohan," A Tabular Approach for Frequent itemset mining." International Journal of Advanced Research in Computer Science, Volume 2, No. 1, Jan-Feb 2011. | [6] H. Ravi Sankar and M.M. Naidu, " A New Approach for Mining Frequent K-itemset," Proceedings of the World Congress on Engineering and Computer Science, San Francisco USA, October 2007. | [7]Christian Borgelt, "Simple Algorithms for Frequent Item Set Mining."European Center for Soft Computing,Asturias, Spain.