A COMPREHENSIVE STUDY OF MAJOR TECHNIQUES OF MULTI LEVEL FREQUENT PATTERN MINING: A SURVEY

Syed Zishan Ali[1], Yogesh Rathore[2]

[1]Computer Science and Engineering, [2]Professor, Computer Science and Engineering, [1, 2]Raipur Institute of Technology, Raipur, Mandir Hasaud, Raipur, Chhattisgarh, INDIA

Abstract

Frequent pattern mining has become one of the most popular data mining approaches for the analysis of purchasing patterns. There are techniques such as Apriority and FP-Growth, which were typically restricted to a single concept level. We extend our research to study Multi - level frequent patterns in multi-level environments. Mining Multi-level frequent pattern may lead to the discovery of mining patterns at different levels of hierarchy. In this study, we describe the main techniques used to solve these problems and give a comprehensive survey of the most influential algorithms
That were proposed during the last decade.

Index Terms: *Data Mining, Data Transformation, Frequent Pattern Mining (FPM), Transactional Database.*

----------***----------

1. INTRODUCTION

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, that appear frequently together in a Transaction data set, is a frequent itemset. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

Frequent pattern mining was first proposed by Agrawal [1] 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 milk, how likely are they going to also buy cereal (and what kind of cereal) 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

One approach to multilevel mining would be to directly exploit the standard algorithms in this area – Apriori [1] and FP-Growth [2] by iteratively applying them in a level by level manner to each concept level. In this paper, we focused on the study of frequent patterns based on the FP- tree [3].

Many scholars have published a tons of research work on frequent pattern mining. There have been extensive studies on the improvements or extensions of Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as PrefixSpan [4],[5], FP-tree [6],[7] have been proposed. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [8]. Some of the basic mining techniques : Apriori, Fp-Growth etc.

2. APRIORI ALGORITHM

Apriori is a algorithm proposed by R. Agrawal and R Srikant in 1994 [1] for mining frequent item sets for Boolean association rule. The name of algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties, as we shall see following. Apriori employs an iterative approach known as level-wise search, where k item set are used to explore (k+1) item sets. There are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step we count the occurrence of each candidate set in database and prunes all disqualified candidates (i.e. all infrequent item sets). Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set The iterations begin with size 2 item sets and the size is incremented after each iteration. The algorithm is based on the closure property of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent.

3. FP-GROWTH ALGORITHM

FP-growth [9] is a well-known algorithm that uses the FP- tree data structure to achieve a condensed representation of the

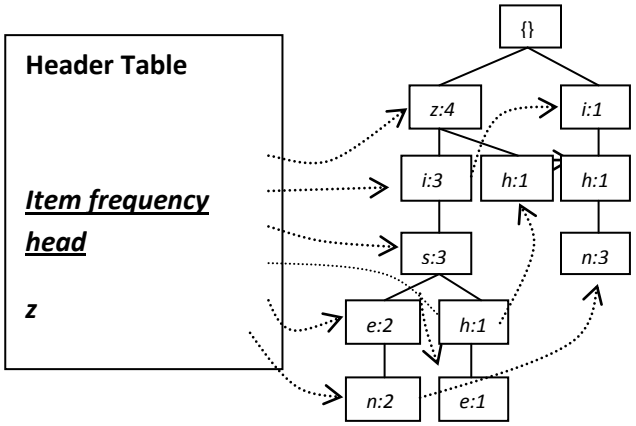
database transactions and employs a divide and-conquer approach to decompose the mining problem into a set of smaller problems. In essence, it mines all the frequent itemsets by recursively finding all frequent itemsets in the conditional pattern base which is efficiently constructed with the help of a node link structure. A variant of FP-growth is the H-mine algorithm [10]. It uses array-based and trie-based data structures to deal with sparse and dense datasets respectively. FPgrowth* [11] uses an array technique to reduce the FP-tree traversal time. In FP-growth based algorithms, recursive construction of the FP-tree affects the algorithm's performance.

TID	*Items bought*	*(ordered) frequent items*
100	**{*z, s, i, d, g, r, e, n*}**	**{*z , i, s, e, n*}**
200	**{*s, h, i, z, l, e, o*}**	**{*z , i, s, h, e*}**
300	**{*h, z, t, j, o, w*}**	**{*z , h*}**
400	**{*h, i , k, s, n*}**	**{*i, h, n*}**
500	**{*s, z, i , e, l ,u, e, n*}**	**{*z , i, s, e, n*}**

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

TABLE 1:

min_support = 3
F-list= z-i-s-h-e-n



4. METHODOLOGY USED

A. ASCENDING FREQUENCY ORDERED PREFIX-TREE (AFOPT)

AFOPT is an efficient algorithm for mining frequent itemsets. It adopts the pattern growth approach and uses a compact data structure---Ascending Frequency Ordered Prefix-Tree (AFOPT) to represent the conditional databases[12]. The AFOPT tree structure is traversed top-down. Compared with the descending frequent order and bottom-up traversal strategy adopted by the FP-growth algorithm, the combination of the top-down traversal strategy and ascending frequency ordering method requires less pointers to be maintained at each node and it also reduces the traversal cost of individual conditional databases.

The goal of mining frequent closed itemsets or maximal frequent itemsets is to reduce output size. An itemset is closed if all of its supersets are less frequent than it. An itemset is maximal if none of its superset is frequent. The complete set of frequent itemsets can be recovered from the set of frequent closed itemsets or the set of maximal frequent itemsets. From frequent closed itemsets, the support information of itemsets can be recovered, but it cannot be recovered from maximal frequent itemsets.

B. ADAPTIVE AFOPT ALGORITHM - ADA AFOPT

An algorithm, named ADA AFOPT, which can be used to resolve the multilevel frequent pattern mining problem. The algorithm is obtained by extending the AFOPT algorithm for multi-level databases. The features of the AFOPT algorithm[12] i.e. FP-Tree, FP-Tree based pattern fragment, 1item partition based divide and conquer method are well preserved. This algorithm uses flexible support constraints. To avoid the problem caused by uniform support threshold, mining with various support constraints is used. Uniform support threshold might cause problems of either generating uninteresting patterns at higher abstraction level or missing potential interesting patterns at lower abstraction are level. At each level, we classify individual items into two categories: normal items and exceptional items.

ADA AFOPT algorithm favours users by pushing these various n transactions, support constraints deep into the mining process. The interestingness of the patterns mound generated, hence, is improved dramatically. Being based on the AFOPT algorithm, this algorithm first traverses the original database to find frequent items or abstract levels and sorts them in ascending frequency order. Then the original database is scanned the second time to construct an AFOPT structure to represent the conditional databases of the frequent items.

C. TRANSACTION REDUCTION TECHNIQUE

Theorem: If c € Fk and c.support < min.support, Titems ≤ k, k = 1, then c is useless in Fk+1 where Fk is Frequent pattern, c is an itemset in each transaction and Titems is total item count in each transactions.

Proof 1: [For c € Fk]
Consider a transaction Ti = {T1, T2, T3… Tm}. Let T1 = {a1, a2, a3… an}, T2 = {a1}. Since c is a hierarchy data. Whenever the lower-level items of c achieve a support count, the higher –level items should be added into the reduced transaction table. For Example data c is 111, 211... Satisfies support count, therefore the higher –level items of 11*, 21*… and the next higher-level also 1**, 2** should be added into the reduced transaction table. If lower-level items of c does not satisfy the min.support, then lower-level items of c is removed from the reduced transaction table. Hence the proof.

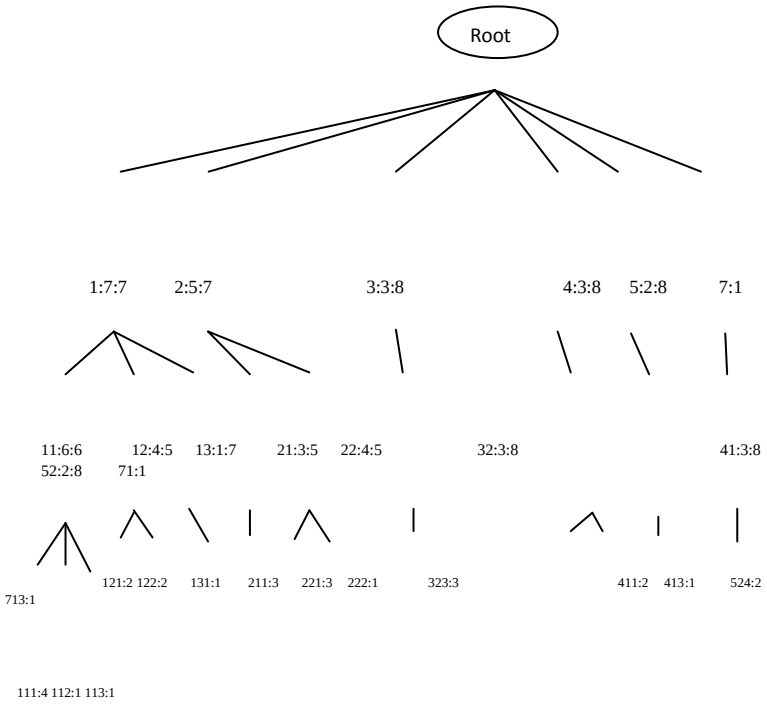
Proof 2: [For T items ≤ k where k = 1]
Now consider the same transaction Ti = {T1, T2, T3… Tm}. Let T1 = {a1, a2, a3… an}, T2 = {a1}. During frequent k+1 pattern generation transaction T2 requires at least 2 as item count and if not then, Ti can be rejected from the transaction table. Hence the proof.

Let us consider the following Example with sample database.

TABLE: 1 Sample Database

TID	Items
T1	{111,121,211,221}
T2	{111,211,222,323}
T3	{112,122,221,411}
T4	{111,121}
T5	{111,122,211,221,413}
T6	{113,323,524}
T7	{131,231}
T8	{323,411,524,713}

CCB – Tree Algorithm [13] has been used to find multilevel frequent 1 pattern.



After generating the FP tree the next step is to generate candidate itemsets and find frequent patterns. It begins by scanning the tree and identifying its leaf nodes. A pointer to each leaf is then inserting into the leaf node array. After that a bottom up scan of each leaf node is done until it reaches the root. Meanwhile each node visited is conserved into temporary buffer for recording the passing path when a node with support count is visited. Candidate Generation keeps the path from starting node i.e. leaf node to the current node and generate all combinations of candidate 2-itemset. Only items from all levels that are above this threshold can be considered as frequent. Candidate itemset which satisfies the minimum support count that candidate can be used for next level processing, the node which does not satisfy minimum .support can be ignored and candidate generation does nothing for this. After finding frequent 2-itemsets from all sub trees. Next traversal is Candidate generation for frequent 3 itemsets. The supports for all the candidate k-itemsets (k≥3) can be computed and the frequent k-itemset can be obtained. This process proceeds until to find frequent k patterns.

TABLE: 2

Techniques	Pre-processing	Feature extraction	Result
Ascending Frequency Ordered Prefix-Tree (AFOPT)	To Traverse the trees in top-down depth-first order	Sorted tree(Ascending order)	Feasible and effective
ADAPTIVE AFOPT algorithm - ADA AFOPT	To Scan the traversed tree	Scanned Tree	Excellent
Transaction Reduction Technique	To Reduce Non candidate Pattern	Reduced Search tree with less I/O cost.	Good

5. COMPARISON

Table 2 shows the comparison between the techniques that are discussed in this paper. The characteristics used to distinguish are: pre-processing, feature extraction, database and result. Pre-processing is the main step in frequent pattern mining.. Second characteristic is feature extraction which gives the extracted features that are used in classification. In AFOPT and ADA-AFOPT the tree is sorted and arranged in ascending order. Transaction Reduction Technique based method is used to reduce the unwanted candidates and transactions and applying the resulted transactions in FP-tree as input to subsequent iterations of the mining process. It reduces the I/O costs and search spaces without losing any patterns.

CONCLUSIONS

This paper's objective is to present the major techniques of multilevel frequent pattern mining. This paper surveys some of the important techniques. The techniques considered in this paper are Ascending Frequency Ordered Prefix-Tree (AFOPT), Adaptive Afopt algorithm (ADA AFOPT), Transaction Reduction Technique. The experimental results shows that AFOPT with ADA AFOPT gives excellent result, Transaction Reduction Technique are also good in minimizing I/O cost.

REFERENCES

[1] Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases.In:Proceedings of the1993 ACM-SIGMOD international conference on management of data (SIGMOD'93), Washington, DC, pp 207–216,.

[2] Han .J ,Pei .J, and Yin .Y,(2000) Mining Frequent patterns without candidate generation. In Proc. Of ACM- SIGMOD Int. Conf. on Management of Data, pages 1- 12.

[3] T.Eavis and XI Zheng, Multi-Level Frequent Pattern Mining, in Springer-Verlag Berlin Heidelberg 2009, pp. 369 – 383..

[4] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," Proc. 17th Int'l Conf. Data Eng. (ICDE '01), pp. 215-224, 2001.

[5] X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2003

[6] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[7] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[8] Y. Li and N. Zhong, "Interpretations of Association Rules by Granular Computing," Proc. IEEE Third Int'l Conf. Data Mining (ICDM '03), pp. 593-596, 2003.

[9] J. Han, J. Pei, and Y. Yin,"Mining Frequent Patterns without Candidate Generation," Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, Dallas, Texas, pp. 1-12, May 2000.

[10] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang, "Hmine: Hyper Structure Mining of Frequent Patterns in Large Databases," Proceedings of IEEE International Conference on Data Mining, pp. 441-448, 2001.

[11] G. Grahne, and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets," FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida,December 2003.

[12] Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. III (2008), Suppl. issue: Proceedings of ICCCC 2008, pp. 437-441

[13] Dr.K.Duraiswamy and B.Jayanthi, a Novel preprocessing Algorithm for Frequent Pattern Mining in Mutidatasets, International Journal of Data Engineering,Vol. 2, No. 3, Aug 2011.

BIOGRAPHIES

Syed Zishan Ali is with the Department of Computer Science and Engineering, hilai Institute of Technology, Raipur, Chhattisgarh, India.. E-mail: zishan786s@gmail.com

Yogesh Rathore is currently the coordinator of M.Tech, and is a Senior lecturer in Department of Computer Science and engineering, Raipur Institute of Technology. Raipur, Chhattisgarh, India. E-mail: yogeshrathore23@gmail.com