RESEARCH ARTICLE OPEN ACCESS

A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth

Siddhrajsinh Solanki[1], Neha Soni[2]
[1]*Computer Engineering Department, SVIT Vasad, India*
[2] *Computer Engineering Department, SVIT Vasad, India*

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

Abstract:

Frequent Pattern Mining is very imporatant task in association mining. Data mining is emerging technology which is continuously increasing its importance in all the aspects of human life. As an important task of data mining, Frequent pattern Mining should understood by researchers to make modification in existing algorithms or to utilize algorithm and methods in more specific way to optimize minig process. This paper concentrate on the study of basic algorithm of frequent pattern mining and its working. It also focus on advantage and disadvantage of algorithms. Basic algorithm studied in this paper are (1) Apriori (2) Eclat (3) FP Growth. Mining of association rules from frequent pattern from massive collection of data is of interest for many industries which can provide guidance in decision making processes such as cross marketing, market basket analysis, promotion assortment etc.

Keywords — Itemset, Frequent Pateern Mining, Apriori, Eclat, Fp Growth.

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

I. INTRODUCTION

T Frequent patterns are the patterns which appear frequently in database. For example a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. If Pattern occurs frequently, it is called a frequent pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Frequent pattern mining has become an important data mining task and a focused theme in data mining research. A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.

Purpose of this paper is to become accustomed to the main important concepts of frequent pattern mining. In data mining we may say that a pattern is a particular data behavior, arrangement or form that might be of a business interest. Itemset is set of items, a group of element that represents together as a single entity.

A frequent itemset is an itemset that occurs frequently .In frequent pattern mining to check whether a itemset occurs frequently or not we have a parameter called support of an itemset . An itemset is termed frequent if its support count is gr eater than the minimum support count set up initially.

I= {i1, i2, i3, …, in} is a set of items, such as products like (computer, CD, printer, papers, …and so on).

Let DB be a set of transactional database where each transaction T is a set of items such that T ⊆ I. Each transaction is associated with unique identifier, transaction identifier (TID).

$$F(D,\sigma)=\{X \subseteq I|\ \text{support} \geqslant \sigma\} \qquad (1.1)$$

The above equation represents that only those items are termed frequent whose support count is greater than the minimum support count initially set up. Association rule is an expression of the from X → Y where X and Y are itemsets and their intersection is null i.e. $X \cap Y=\{\}$.

The support of an association rule is the support of the union of X and Y, i.e. X is called the head or antecedent and Y is called the tail or consequent of the rule.

The confidence of an association rule is defined as the percentage of rows in D containing itemset X that also contain itemset Y, i.e.

$$\text{CONFIDENCE}(X \rightarrow Y)$$
$$P(X|Y)= \text{SUPPORT}(X \cup Y)/\text{SUPPORT}(X) \qquad (1.2)$$

Association Mining is Two-step approach:

Frequent Itemset Generation

– Generate all itemsets whose support > minsup

Rule Generation

– Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

Once frequent itemsets or pattern are mine then it can be easy to find association between them.

II. TECHNIQUES FOR FREQUENT PATTERN MINING

There are various techniques are proposed for generating frequent itemsets so that association rules are mined efficiently. The approaches of generating frequent itemsets are divided into basic three techniques.

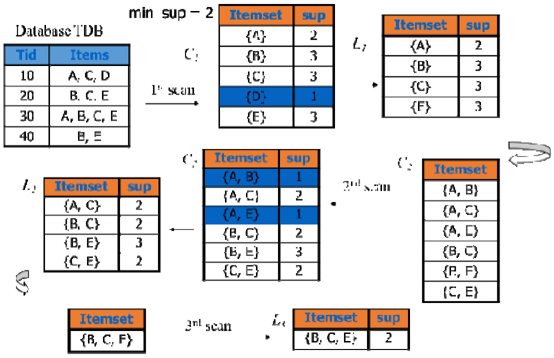
1. Apriori Algorithm : Horizontal Layout based
2. Eclat Algorithm : Vertical Layout based
3. FP Growth Algorithm : Projected database based.

III. APRIORI ALGORITHM

Apriori[2] is the most classical and important algorithm for mining frequent itemsets. Apriori is used to find all frequent itemsets in a given database DB. The key idea of Apriori algorithm is to make multiple passes over the database. Apriori algorithm fairly depends on the apriori property which states that "All non empty itemsets of a frequent itemset must be frequent"[2]. It also described the anti monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test [2, 3].

Apriori algorithm follows two phases:

- Generate Phase: In this phase candidate (k+1)-itemset is generated using k-itemset,this phase creates Ck candidate set.
- Prune Phase: In this phase candidate set is pruned to generate large frequent itemset using "minimum support" as the pruning parameter.This phase creates Lk large itemsetse.



Tid	Items
10	A, C, D
20	B. C. E
30	A, B, C, E
40	B, E

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Itemset	sup
{A}	2
{B}	3
{C}	3
{F}	3

Itemset
{A, B}
{A, C}
{A, D}
{B, C}
{P, F}
{C, E}

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

Itemset
{B, C, F}

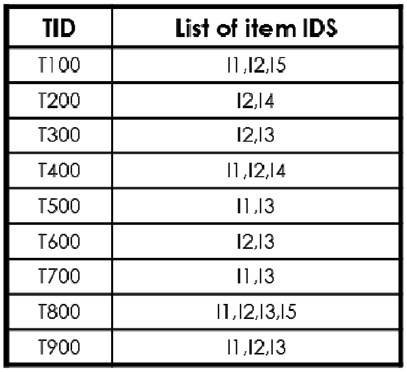
Itemset	sup
{B, C, E}	2

Figure 1 Example of Apriori

Disadvantage: It need to generate a huge number of candidate sets. It need to repeatedly scan the database and check a large set of candidates by pattern matching. It is costly to go over each transaction in the database to determine the support of the candidate itemsets.

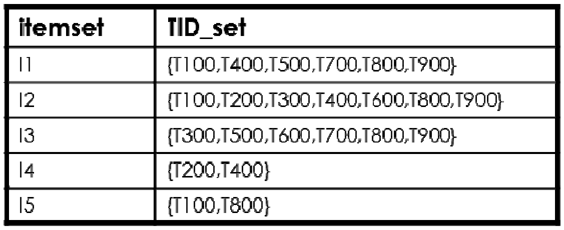
IV. ECLAT ALGORITHM

Eclat[4] algorithm is a depth first search based algorithm. It uses a vertical database layout i.e. instead of explicitly listing all transactions; each item is stored together with its cover (also called tidlist) and uses the intersection based approach to compute the support of an itemset [4].It requires less space than apriori if itemsets are small in number .It is suitable for small datasets and requires less time for frequent pattern generation than apriori. Below is the Example of Eclat Algorithm for minimum support = 2.



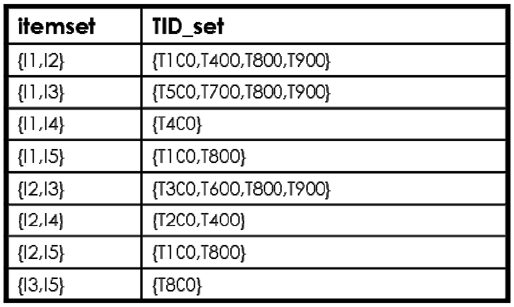
TID	List of item IDS
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Fig. 2 Database in Horizontal Data Format



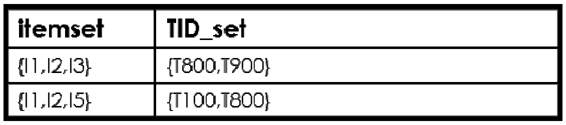
itemset	TID_set
I1	{T100,T400,T500,T700,T800,T900}
I2	{T100,T200,T300,T400,T600,T800,T900}
I3	{T300,T500,T600,T700,T800,T900}
I4	{T200,T400}
I5	{T100,T800}

Fig. 3 Database in Vertical Data Format



itemset	TID_set
{I1,I2}	{T1C0,T400,T800,T900}
{I1,I3}	{T5C0,T700,T800,T900}
{I1,I4}	{T4C0}
{I1,I5}	{T1C0,T800}
{I2,I3}	{T3C0,T600,T800,T900}
{I2,I4}	{T2C0,T400}
{I2,I5}	{T1C0,T800}
{I3,I5}	{T8C0}

Fig. 4 2-Itemset generated by intersection of 1-itemset



itemset	TID_set
{I1,I2,I3}	{T800,T900}
{I1,I2,I5}	{T100,T800}

Fig. 5 Frequent 3-itemset generated by intersection of 2-itemset

Disadvantages: Whe tid-list is large at that time it takes more space to store candidate set. It needs more time for intersection when Tid list is large.

V. FP GROWTH ALGORITHM

FP Growth[7] is another important frequent pattern mining method, which generates frequent itemset without candidate generation. It uses tree based structure. The problem of Apriori algorithm was dealt with, by introducing a novel, compact da ta structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed[7]. It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum support[2].FP-growth traces the set of concurrent items[7].

FP tree is constructed in two passes:

Pass 1:

- Scan data and count support for each item
- Discard infrequent items
- Sort frequent items in descending order based on their support

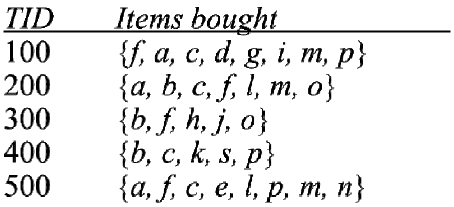
Pass 2:

- Reads one transaction at a time and maps it to the tree
- Fixed order is used so that path can be shared
- Pointers are maintained between nodes containing same items
- Frequent items are exctracted from the list

It suffers from certain disadvantages:

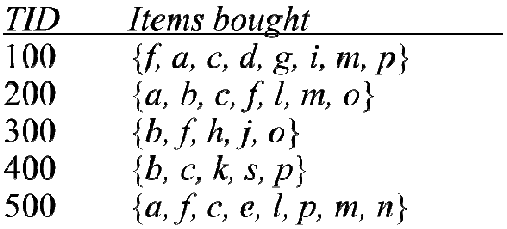
- Fp tree may not fit in main memory
- Execution time is large due to complex compact data structure[5]

Below is Example FP Growth Algorithm.



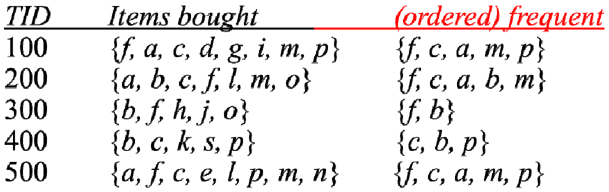
TID	*Items bought*
100	{*f, a, c, d, g, i, m, p*}
200	{*a, b, c, f, l, m, o*}
300	{*b, f, h, j, o*}
400	{*b, c, k, s, p*}
500	{*a, f, c, e, l, p, m, n*}

Fig. 6 Database



TID	*Items bought*
100	{*f, a, c, d, g, i, m, p*}
200	{*a, b, c, f, l, m, o*}
300	{*b, f, h, j, o*}
400	{*b, c, k, s, p*}
500	{*a, f, c, e, l, p, m, n*}

Fig. 7 L after scanning first time



TID	Items bought	(ordered) frequent
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Fig. 8 Ordered Frequent Itemset

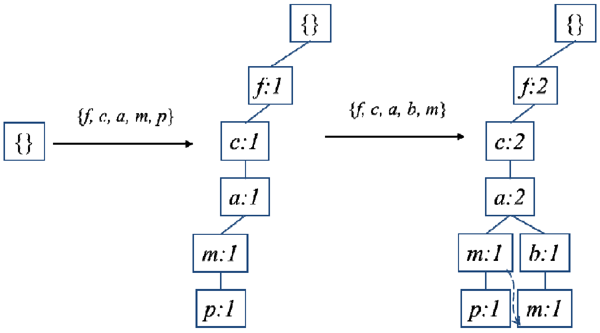


Fig. 9 Construct FP tree (a)

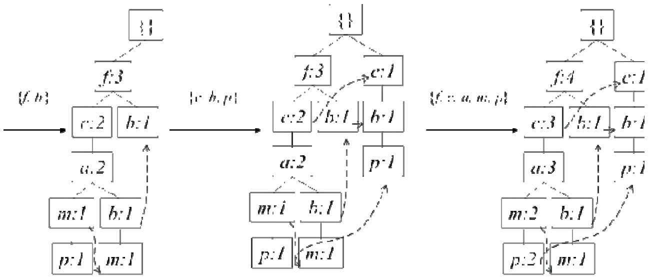


Fig. 10 Construct FP Tree(b)

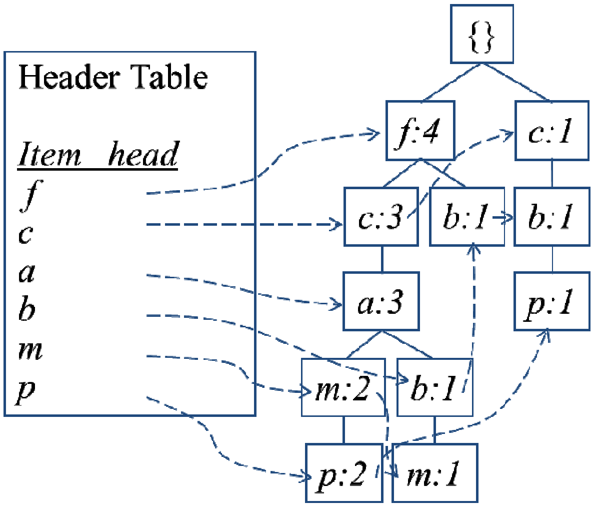


Fig. 11 Final FP Tree

VI. CONCLUSION

Frequent Pattern Mining is very imporatant step of association mining. We have discussed three basic algorithm of frequent pattern mining (1) Apriori (2)Eclat (3)FP Growth. All three are of very imporatant. We have discussed their advantages and disadvantages. So Apriori needs more database scan. Eclat needs one database scan and it find next level itemsets by intersecting current level itemsets. FP Growth is taking advantage of its tree structure. But it uses complex data structure compare Aprori and Eclat.

ACKNOWLEDGMENT

We express our gratitude to management of SVIT, Vasad for proving opportunity and support for such an activity.

REFERENCES

[1] Dr. Kanwal Garg Deepak Kumar “ Comparing the Performance of Frequent Pattern Mining Algorithms” International Journal of Computer Applications (0975 – 8887) Volume 69– No.25, May 2013

[2] Agrawal.R and Srikant.R. “Fast algorithms for mining association rules”. In Proc. Int’l Conf. Very Large Data Bases (VLDB), pages 487–499, Sept. 1994.

[3] Bart Goethals,“Survey on Frequent Pattern Mining”, HIIT Basic Research Unit,Department of Computer Science,University of Helsinki,Finland.

[4] M. J. Zaki, S. ParthaSarathy, M. Ogihara, W. Li “New Algorithms for Fast Discovery of Association Rules” KDD 97

[5] C.Borgelt. “Efficient Implementations of Apriori and Eclat”. In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 2003.

[6] Mohammed J. Zaki and Karam Goudaz “Fast Vertical Mining Using Diffsets” KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining ACM New York, NY, USA ©2003

[7] Han J., Pei H., and Yin. Y., Mining Frequent Patterns without Candidate Generation, In Proc. Conf. on the Management of Data (2000)

[8] Pei. J, Han. J, Lu. H, Nishio. S. Tang. S. and Yang. D., Hmine: Hyper-structure mining of frequent patterns in large databases, In Proc. Int’l Conf. Data Mining (2001)

[9] Pratiksha Shendge ,Tina Gupta, “Comparitive Study of Apriori & FP Growth Algorithms”, PARIPEX - INDIAN JOURNAL OF RESEARCH ISSN - 2250-1991 Volume : 2 | Issue : 3 | March 2013