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

[1]Aakansha Saxena, [2]Sohil Gadhiya
[1]PG Student, [2]Assitant professor,
[1,2]Computer engineering, C.U. Shah College of Engineering and Technology,Wadhwan, India.
[1]aakanshasaxena_83@yahoo.com, [2]sohilgadhiya@gmail.com

***Abstract*— Frequent pattern mining is one of the most important task for discovering useful meaningful patterns from large collection of data.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. The techniques of discovering association rule from data have traditionally focused on identifying relationship between items predicting some aspect of human behavior, usually buying behavior. In this paper ,the study includes three classical frequent pattern mining methods that are Apriori, Eclat, FP growth and discusses some issues related with these algorithms.**

***Keywords*— Frequent pattern mining, Apriori,FP growth, Eclat**

I. Introduction

In recent years amount of data in the database has increased rapidly. The increasing size of the database has led to growing interest in extraction of useful information from the bulk of data. Data mining is a technique useful for attaining useful information from vast databases. Implicit information within a database can be very useful in tasks such as marketing, financial forecast etc. These information have to be derived efficiently.Frequent pattern mining discovers significant relationships among variables or items in a dataset.

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 occurrs 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 greater 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 $T \subseteq I$. Each transaction is associated with unique identifier, transaction identifier (TID).

$$F(D,\sigma)=\{X \subseteq I \mid \text{support} \geq \sigma\} \qquad (1)$$

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 $X\rightarrow Y$ 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.

$$\textsc{Confidence}(X\rightarrow Y)=P(X|Y)=\textsc{support}(X\cup Y)/\textsc{support}(X) \qquad (2)$$

II. Techniques for frequent pattern mining

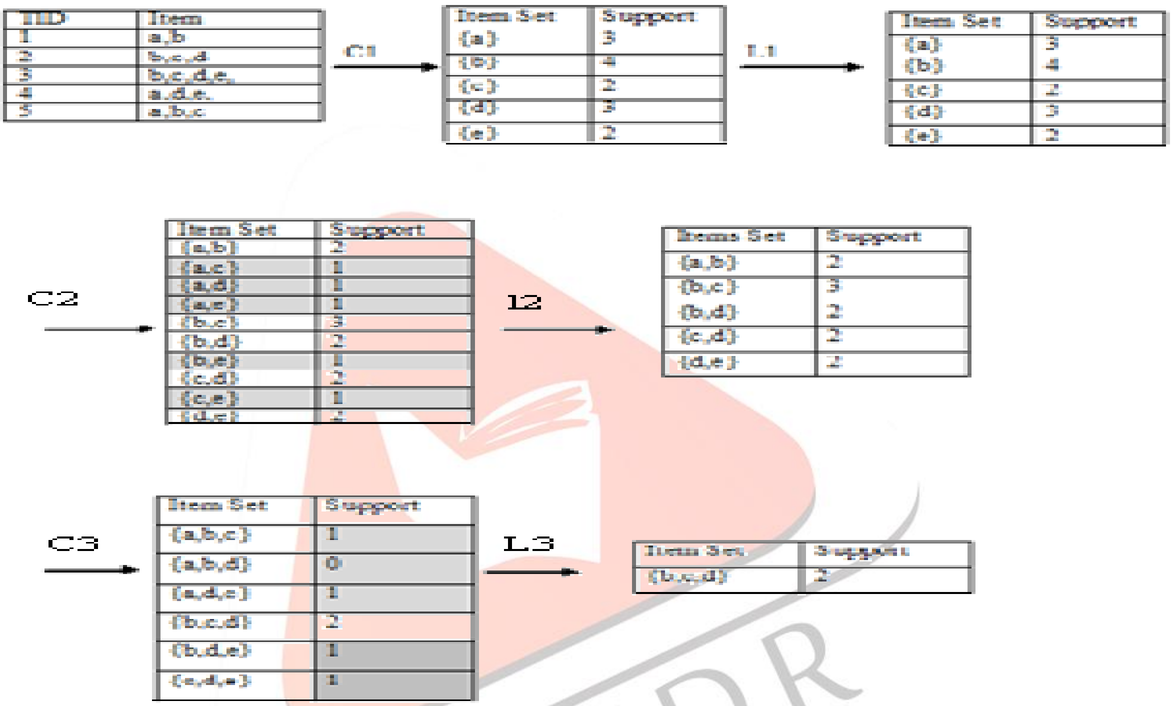
There are various techniques are proposed for generating frequent itemsets so that association rules are mined efficiently. The approaches of generating frequent itemsets are divided into basic three techniques.

- Horizontal layout based data mining techniques : Apriori algorithm
- Vertical layout based data mining techniques : Eclat algorithm
- Projected database based data mining techniques : FP- Growth algorithm

III. APRIORI ALGORITHM

This 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 C_k 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 L_k large itemset



TID	Item
1	a,b
2	b,c,d
3	b,c,d,e,
4	a,d,e.
5	a,b,c

Item Set	Support
{a}	3
{b}	4
{c}	2
{d}	3
{e}	2

Item Set	Support
{a}	3
{b}	4
{c}	2
{d}	3
{e}	2

Item Set	Support
{a,b}	2
{a,c}	1
{a,d}	1
{a,e}	1
{b,c}	3
{b,d}	2
{b,e}	1
{c,d}	2
{c,e}	1
{d,e}	2

Items Set	Support
{a,b}	2
{b,c }	3
{b,d}	2
{c,d}	2
{d,e }	2

Item Set	Support
{a,b,c}	1
{a,b,d}	0
{a,d,e}	1
{b,c,d}	2
{b,d,e}	1
{c,d,e}	1

Item Set	Support
{b,c,d}	2

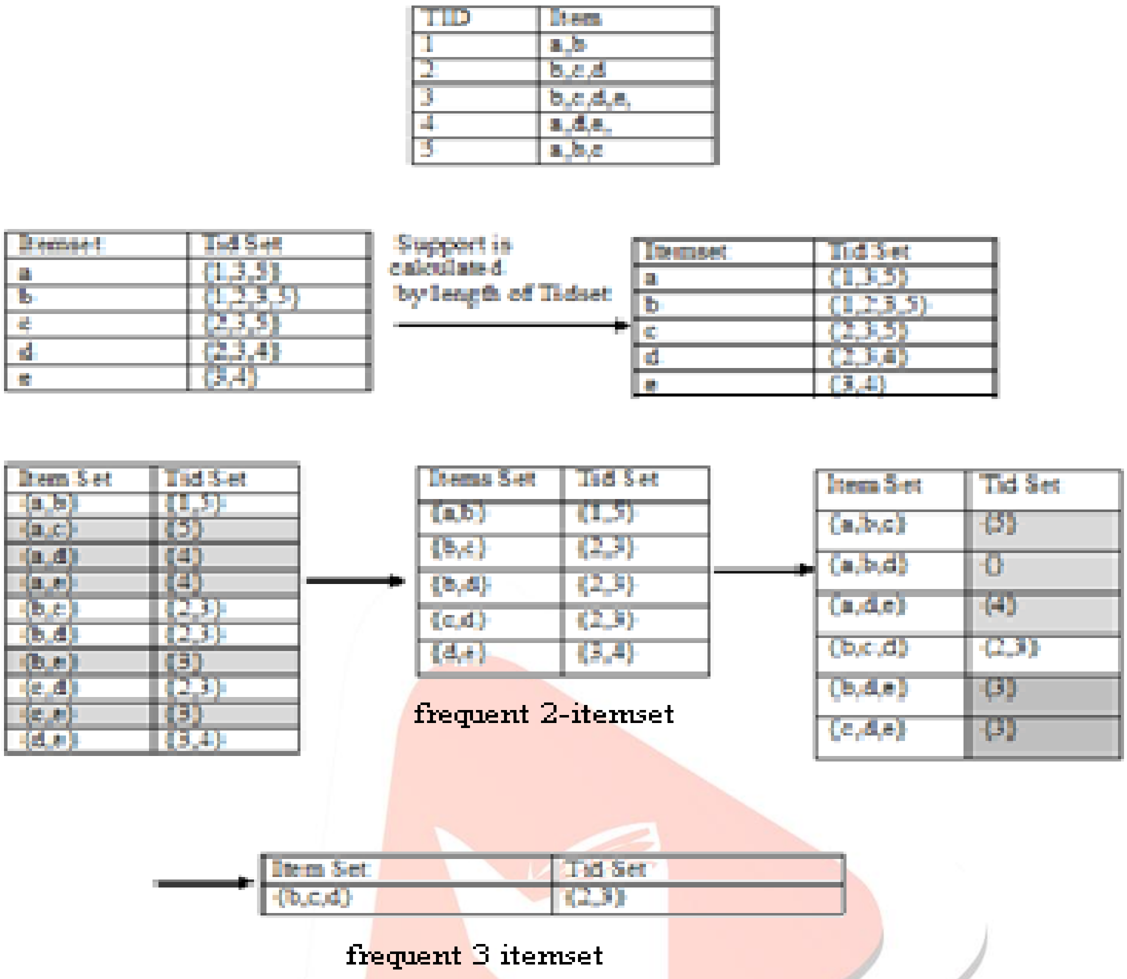
Fig. 2 : Example of Apriori algorithm

These disadvantages can be minimized by applying techniques to:

- Reduce passes of transaction database scans
- Shrink number of candidates
- Facilitate support counting of candidates

IV. ECLAT ALGORITHM

Eclat 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 [5].It requires less space than apriori if itemsets are small in number [5].It is suitable for small datasets and requires less time for frequent pattern generation than apriori.



TID	Item
1	a,b
2	b,c,d
3	b,c,d,e,
4	a,d,e,
5	a,b,c

Itemset	Tid Set
a	{1,3,5}
b	{1,2,3,5}
c	{2,3,5}
d	{2,3,4}
e	{3,4}

Support is calculated by length of Tidset

Itemset	Tid Set
a	{1,3,5}
b	{1,2,3,5}
c	{2,3,5}
d	{2,3,4}
e	{3,4}

Item Set	Tid Set
{a,b}	{1,5}
{a,c}	{5}
{a,d}	{4}
{a,e}	{4}
{b,c}	{2,3}
{b,d}	{2,3}
{b,e}	{3}
{c,d}	{2,3}
{c,e}	{3}
{d,e}	{3,4}

Items Set	Tid Set
{a,b}	{1,5}
{b,c}	{2,3}
{b,d}	{2,3}
{c,d}	{2,3}
{d,e}	{3,4}

frequent 2-itemset

Item Set	Tid Set
{a,b,c}	{5}
{a,b,d}	{}
{a,d,e}	{4}
{b,c,d}	{2,3}
{b,d,e}	{3}
{c,d,e}	{3}

Item Set	Tid Set
{b,c,d}	{2,3}

frequent 3 itemset

Fig 2:Example of eclat algorithm

V. FP GROWTH

This 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 data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed[5]. It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum support[5].FP-growth traces the set of concurrent items[6].

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

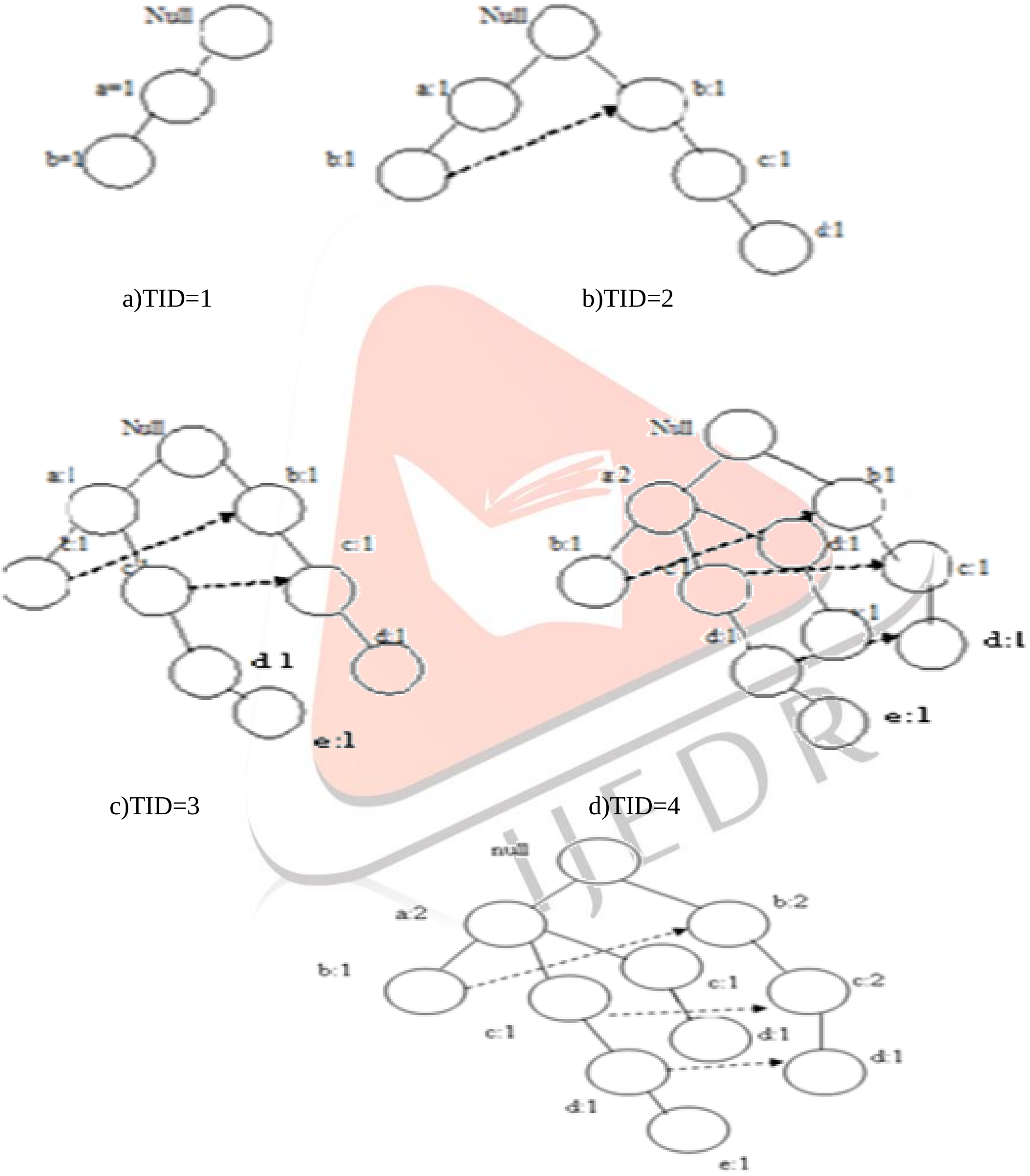
Pass 2:

- 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

TID	Item
1	a,b
2	b,c,d

3	b,c,d,e,
4	a,d,e,
5	a,b,c

Table 1).Transaction Table



a)TID=1

b)TID=2

c)TID=3

d)TID=4

e)TID=5

Fig. c) Fp tree construction from transactions in Table 1.

It suffers from certain disadvantages:

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

VI. CONCLUSION

Frequent pattern mining is an important task in association rule mining. It has been found useful in many application like market basket analysis, financial forecasting etc.We have discussed about three classical algorithm Apriori,Fp growth and eclat with their pros and cons. using the horizontal approach ,owing to all candidate itemset for each level has to be discovered ,the longer the length of the frequent itemset ,more the number of candidate generation. Projected tree method is efficient in terms of speed but utilizes more space. These disadvantages can be overcome by using techniques like hashing, partitioning etc.

REFERENCES

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

[2]Aggrawal.R, Imielinski.t, Swami.A. "Mining Association Rules between Sets of Items in Large Databases". In Proc. Int'l Conf. of the 1993 ACM SIGMOD Conference Washington DC, USA.

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

[4]Renáta Iváncsy, István Vajk, "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[5]Pramod S.,O.P. Vyas "Survey on Frequent Item set Mining Algorithms", International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 15

[6]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

[7]Mona S Kamat,J.W. Bakal,Madhu nashipudi, "Comparitive study techniques to Discover frequent pattern of web usage mining", International Journal on Advanced Computer Theory and Engineering (IJACTE) ISSN

[8]Sachin Sharma, Vidushi Singhal and Seema Sharma, "A SYSTEMATIC APPROACH AND ALGORITHM FOR FREQUENT DATA ITEMSETS", Journal of Global research in computer science, Volume 3, No. 11, November 2012