UNIT-II

Association rule mining

⇒ Association rule mining is a a step process.

1) finding frequent patterns
2) forming strong association rules. → For this we use pattern evaluation methods.

& Algorithms

1) Apriar : Algorithm
2) Fp growth Algorithm.

Apriar Algorithm:- R. Agarwal and R. Srikanth developed the algorithm in 1994.

⇒ Main purpose of the algorithm is to find frequent pattern and forming boolean rules.

⇒ This algorithm uses prior knowledge of item set properties

⇒ It uses level wise search

⇒ It is a a step process

1) join step
2) prune step.

Step 1:- Join step :- To find L_k a set of candidate k-items [L_k(k-item)] is generated by joining L_{k-1} with itself. This set of candidates is denoted by C_k

Step 2:- Prune step :- C_k is a superset of L_k i.e its members may (or) may not be frequent but all of the

K-frequent item sets are included in C_k to reduce the size of C_k. The Apriori property is used.

Eg:- By using apriori algorithm, find the items that are frequently purchased by the customers

TID	List of item - IDs
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

Minimum support count = 2

Scan 'D' for count of each candidate

C_1 (candidate itemset 1)

Itemset	Suppcount
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

compare Candidate support count with minimum support count

L_1

Itemset	Support count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

Generate C_2 candidates from L_1

C_2 (candidates set 2)

Itemset
$\{I_1, I_2\}$
$\{I_1, I_3\}$
$\{I_1, I_4\}$
$\{I_1, I_5\}$
$\{I_2, I_3\}$
$\{I_2, I_4\}$
$\{I_2, I_5\}$
$\{I_3, I_4\}$
$\{I_3, I_5\}$
$\{I_4, I_5\}$

Scan 'D' for count of each candidate

Itemset	Support count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

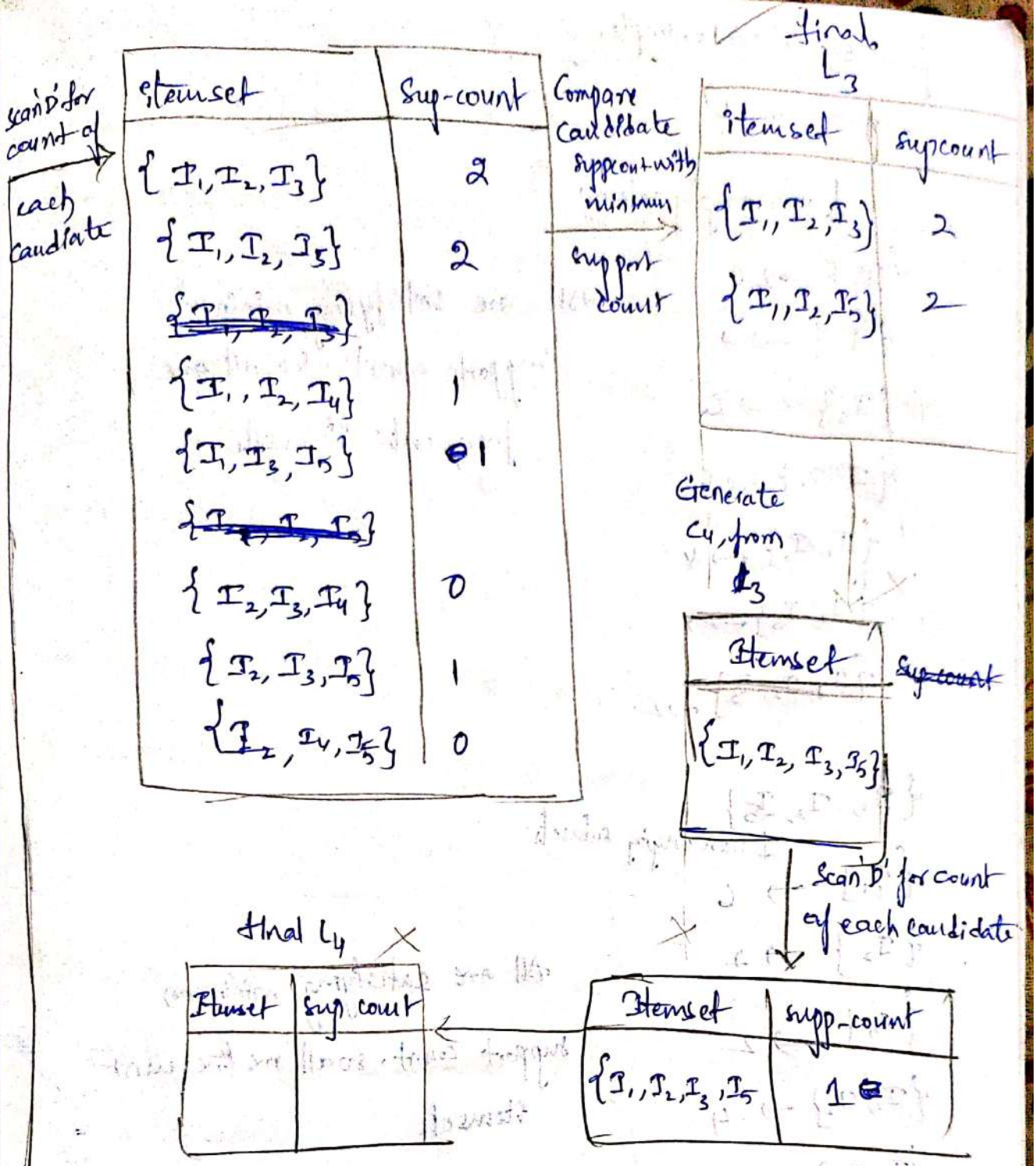
Find L_2 ↓ Compare candidate support count with minimum sup count

Itemset	Sup count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Generate C_3 Candidates from L_2

C_3

Itemset
$\{I_1, I_2, I_3\}$
$\{I_1, I_2, I_5\}$
~~$\{I_1, I_2, I_3\}$~~
$\{I_1, I_2, I_4\}$
$\{I_1, I_3, I_5\}$
~~$\{I_1, I_4, I_5\}$~~
$\{I_2, I_3, I_4\}$
$\{I_2, I_3, I_5\}$
$\{I_2, I_4, I_5\}$



∴ frequent item sets are. $\{I_1, I_2, I_3\}$ and $\{I_1, I_2, I_5\}$

Apriori property :-

All non empty subsets of frequent item sets must also be frequent

from the above example.

$\{I_1, I_2, I_3\}$

↓ non empty subsets

$\{I_1\} \rightarrow 6$

$\{I_2\} \rightarrow 7$

$\{I_3\} \rightarrow 6$

$\{I_1, I_2\} \rightarrow 4$

$\{I_1, I_3\} \rightarrow 4$

$\{I_2, I_3\} \rightarrow 4$

$\{I_1, I_2, I_3\} \rightarrow 2$

All are satisfying minimum support count, so all are frequent itemsets.

$\{I_1, I_2, I_5\}$

↓ non empty subsets

$\{I_1\} \rightarrow 6$

$\{I_2\} \rightarrow 7$

$\{I_5\} \rightarrow 2$

$\{I_1, I_2\} \rightarrow 4$

$\{I_1, I_5\} \rightarrow 2$

$\{I_2, I_5\} \rightarrow 2$

$\{I_1, I_2, I_5\} \rightarrow 2$

All are satisfying minimum support count, so all are frequent itemsets

Generate strong Association Rules from frequent Itemset (10m) :-

Association rule :- Eg :- 1) milk ⇒ Bread (antecedent → consequent)

2) buys (onion, potatoes) ⇒ buys (Tomatoes) (antecedent → consequent)

Def :- Association rules are if-then statements. They

It is used for analysing and predicting customer behaviour.

⇒ In data mining association rules are useful in uncovering relationship btw unrelated data.

⇒ Association rules contain 2 parts

1) Antecedents

2) Consequent

⇒ Antecedent :- can be found alone

⇒ Consequent will be found with combination of antecedent

Eg 3 :- computer ⇒ Antivirous software [support = 2%, confidence = 60%]

⇒ A support of 2% for above rule means that 2% of all transaction under analysis shows that computer and antivirous s/w are purchased together.

⇒ A Confidence of 60% means that 60% of customers who purchased a computer also purchased antivirous software

⇒ Typically association rules are considered interesting (or strong if they satisfy both a minimum support threshold and minimum confidence threshold.

⇒ These thresholds can be set by users. (or domain experts.

Generating strong Association rules from Frequent Itemset (ioM):-

$$\text{Support}(A \Rightarrow B) = P(A \cup B)$$

$$\text{Confidence}(A \Rightarrow B) = P(B/A)$$

$$= \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

$$= \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

⇒ support count of $A \cup B$ = no. of transactions containing the items A and B.

⇒ support count of A = no. of transactions containing the item A.

⇒ we can find strong association rules by using the confidence

⇒ for each frequent itemset 'l' generate all non empty subsets of 'l'.

⇒ for every non-empty subset 's' of 'l' output the rule "$s \Rightarrow (l-s)$" if support count of 'l'

$$\text{if } \frac{\text{support count of } l}{\text{support count of } s} \geq \text{minimum confidence.}$$

Eg :- $\{I_1, I_2, I_5\} \rightarrow$ 'l'

↓ non empty subsets

"$s \Rightarrow (l-s)$" Confidence = support count $\{I_1 \cup I_2, I_5\}$ / support count $\{I_1\}$

R_1	's' $\rightarrow \{I_1\}$	$\longrightarrow \{I_2, I_5\}$	$\Rightarrow 2/6 * 100 = 33.33\%$
R_2	$\{I_2\}$	$\longrightarrow \{I_1, I_5\}$	$\Rightarrow 2/7 \times 100 = 28.57\%$
R_3	$\{I_5\}$	$\longrightarrow \{I_1, I_2\}$	$\Rightarrow 2/2 * 100 = 100\%$
R_4	$\{I_1, I_2\}$	$\longrightarrow \{I_5\}$	$\Rightarrow 2/4 * 100 = 50\%$
R_5	$\{I_1, I_5\}$	$\longrightarrow \{I_2\}$	$\Rightarrow 2/2 * 100 = 100\%$
R_6	$\{I_2, I_5\}$	$\longrightarrow \{I_1\}$	$\Rightarrow 2/2 \times 100 = 100\%$
R_7	$\{I_1, I_2, I_5\}$	$\longrightarrow \{\}$ (or) ϕ	

minimum confidence = 60%

R_5, R_6, R_3 are strong association rules.

FR - Growth Algorithm :- It is a 2 step process.

1) FP - tree
2) Conditional databases

FP - tree construction / Generation :-

step 1 :- Write list of itemsets with support count

step 2 :- Write the items in descending order of their support-count.

step 3 :- Start construction of FP tree with null as root node (Arrange items of every transaction in descending order of supp-count)

step 4 :- Link to nodes.

Eg :-

T ID	List of items Bo's
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

minimum support count = 2

min support confidence = 60%

step 1:-

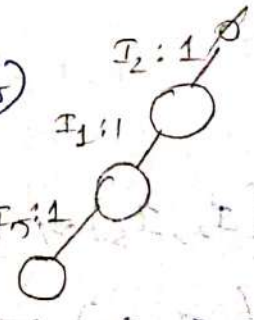
Itemset	support count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

step 2:-

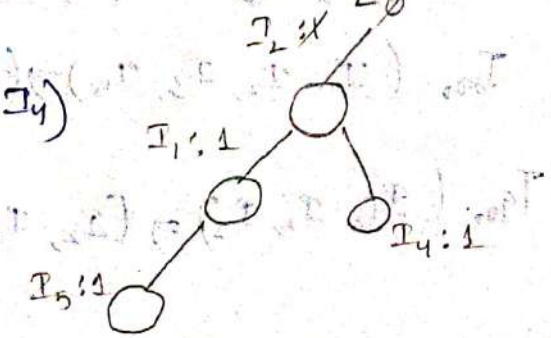
Itemset	support count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

step 3:- $T_{100}(I_1, I_2, I_5) \Rightarrow (I_2, I_1, I_5)$

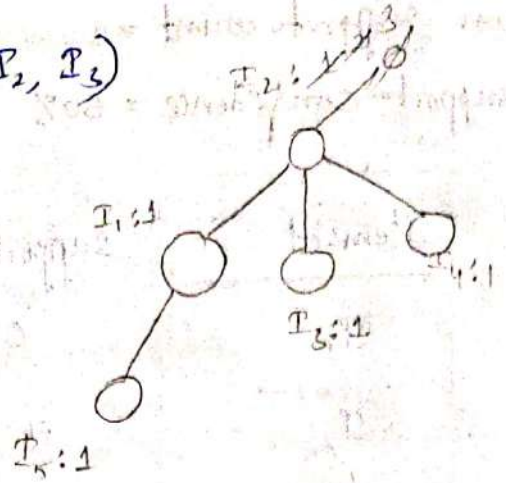
(start every transaction from ϕ)



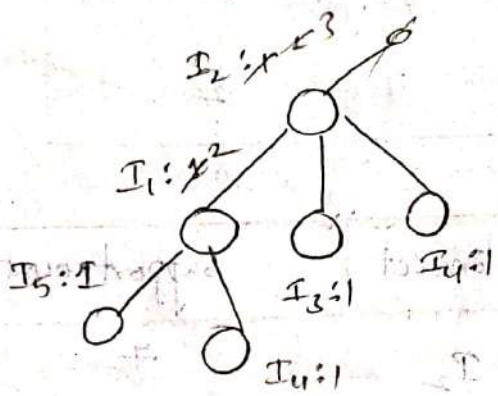
$T_{200}(I_2, I_4) \Rightarrow (I_2, I_4)$



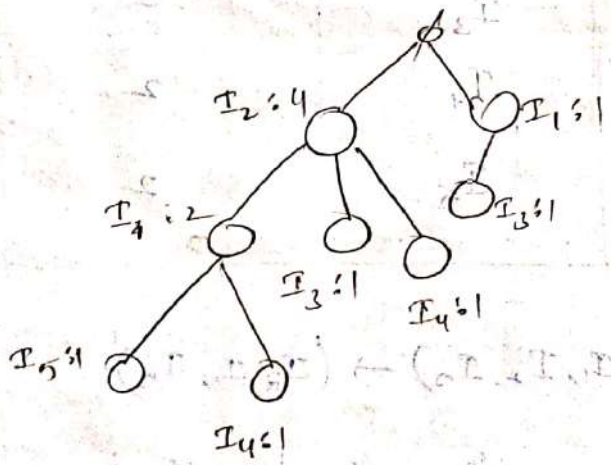
$T_{300}(I_2, I_3) \Rightarrow (I_2, I_3)$



$T_{400}(I_1, I_2, I_4) = (I_2, I_1, I_4)$



$T_{500}(I_1, I_3) \Rightarrow (I_1, I_3)$

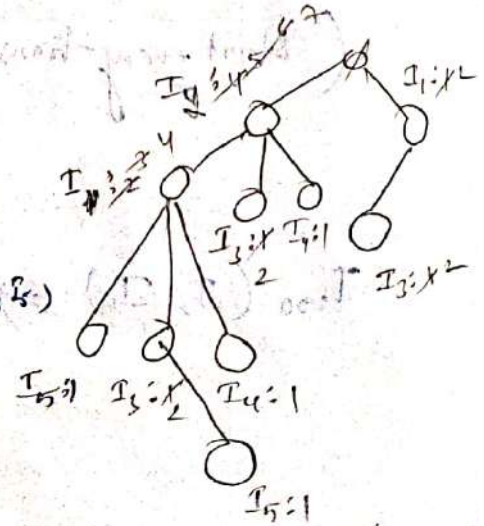


$T_{600}(I_2, I_3) \Rightarrow (I_2, I_3)$

$T_{700}(I_1, I_3) \Rightarrow (I_1, I_3)$

$T_{800}(I_1, I_2, I_3, I_5) \Rightarrow (I_2, I_1, I_3, I_5)$

$T_{900}(I_1, I_2, I_3) \Rightarrow (I_2, I_1, I_3)$



Step 4:-

Item ID	support count	Node link
I_2	7	
I_1	6	
I_3	6	
I_4	2	
I_5	2	

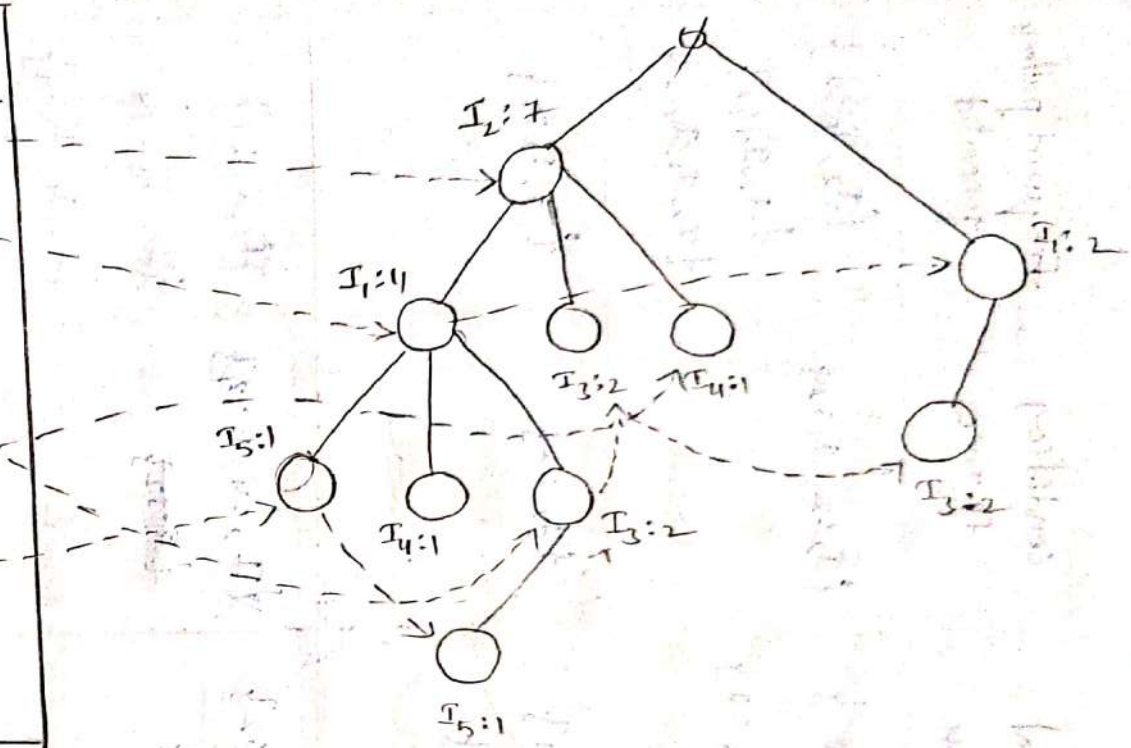


Fig:- An FI-Tree registers compressed frequent patterns information ..

Conditional databases:-

Item	Conditional pattern base (path in the tree)	Conditional FP-Tree (in the tree all paths)	Frequent patterns generated.
I_5 (least sup. count)	$\{I_2, I_1 : 1\}$, $\{I_2, I_1, I_3 : 1\}$ ($I_3 : 1 < 2$) X (1 = count a node)	$\{I_2 : 2, I_1 : 2\}$	$\{I_2, I_5 : 2\}$, $\{I_1, I_5 : 2\}$, $\{I_2, I_1, I_5 : 2\}$
I_4 (next least sup. count)	$\{I_2, I_1 : 1\}$, $\{I_2 : 1\}$, $I_1 : 1 < 2$ X	$\{I_2 : 2\}$	$\{I_4, I_2 : 2\}$
I_3	$\{I_2, I_1 : 2\}$ (left), $\{I_2 : 2\}$, $\{I_1 : 2\}$ (right to ϕ)	$\{I_2 : 4, I_1 : 4\}$ ~~$\{I_1 : I\}$~~	$\{I_2, I_3 : 4\}$, $\{I_1, I_3 : 4\}$, $\{I_2, I_1, I_3 : 4\}$
I_1	$\{I_2 : 4\}$ X (no intermediate node)	$\{I_2 : 4\}$	$\{I_1, I_2 : 4\}$
I_2	No intermediate node		

for a given transactional data set generate rules by using apriar algorithm. Consider the values as support = 50%, confidence = 75%.

Transaction ID	Items purchased.
1	Bread, cheese, egg, juice
2	Bread, cheese, juice.
3	Bread, milk, yogurt
4.	Bread, juice, milk
5	cheese, juice, milk.

Item	frequency	support
Bread	4	$\frac{4}{5} \times 100 = 80\%$
cheese	3	$\frac{3}{5} \times 100 = 60\%$
egg	1	$\frac{1}{5} \times 100 = 20\%$
juice	4	$\frac{4}{5} \times 100 = 80\%$
milk	3	$\frac{3}{5} \times 100 = 60\%$
yogurt	1	$\frac{1}{5} \times 100 = 20\%$

egg, yogurt not have the support value i.e 50% to remove both, i.e support of egg and yogurt is less than 50%.

Item	Frequency	Support
{Bread, cheese}	2	$\frac{2}{5} \times 100 = 40\%$ ✗
{Bread, juice}	3	$\frac{3}{5} \times 100 = 60\%$ ✓
{Bread, milk}	2	$\frac{2}{5} \times 100 = 40\%$ ✗
{cheese, juice}	3	$\frac{3}{5} \times 100 = 60\%$ ✓
{cheese, milk}	1	$\frac{1}{5} \times 100 = 20\%$ ✗
{juice, milk}	2	$\frac{2}{5} \times 100 = 40\%$ ✗

only {Bread, juice} and {cheen, juice} satisifies the support count. so consider only these two sets and remove the remaining sets.

Item	Frequency	support
{Bread, juice, cheese}	2	$\frac{2}{5} \times 100 = 40\%$

As the set {Bread, juice, cheen} donot satisfy the support count. so consider the previous table frequencies which satisfy the support i.e {Bread, juice} and {cheen, juice}

1. {Bread, juice}

Rule 1:- Bread ⇒ juice → Confidence

$$= \frac{\text{Support count}(B \cup J)}{\text{Support count}(B)}$$

$$= \frac{3/5}{4/5} = \frac{3}{4} = 75\%$$

Rule 2:- Juice → Bread = $\frac{\text{Support count } (J \cup B)}{\text{Support } (J)}$

$= \frac{3/5}{4/5}$

$= 75\%$

2) {cheese, juice} Confidence

Rule 3:- {cheese} ⇒ juice $= \frac{3/5}{3/5} = 100\%$

Rule 4: juice ⇒ cheese $= \frac{3/5}{4/5} = 75\%$

All Rules i.e Rules 1, 2, 3, 4. are satisfying the minimum confidence threshold i.e 75%

Hence all rules are strong association rules.

Eg:- 3

For the following given transaction data set generate rules using apriar algorithm. Consider values as support = 22% and confidence = 70%

Transaction ID	Items purchased.
1	I_1, I_2, I_5
2	I_2, I_4
3	I_2, I_3
4	I_1, I_2, I_4
5	I_1, I_3
6	I_2, I_3
7	I_1, I_3
8	I_1, I_2, I_3, I_5
9	I_1, I_2, I_3

Item	Frequency	Support
I_1	6	$\frac{6}{9} \times 100 = 66.6 = 66$
I_2	7	$\frac{7}{9} \times 100 = 77.7 = 77$
I_3	6	$\frac{6}{9} \times 100 = 66.6 = 66$
I_4	2	$\frac{2}{9} \times 100 = 22.2 = 22$
I_5	2	$\frac{2}{9} \times 100 = 22.2 = 22$

All items are ~~support~~ satisfying support count.

Item	Frequency	Support
$\{I_1, I_2\}$	4	$\frac{4}{9} \times 100 = 44$
$\{I_1, I_3\}$	4	$\frac{4}{9} \times 100 = 44$
$\{I_1, I_4\}$	1	$\frac{1}{9} \times 100 = 11.11 \times$
$\{I_1, I_5\}$	2	$\frac{2}{9} \times 100 = 22$
$\{I_2, I_3\}$	4	$\frac{4}{9} = 44$
$\{I_2, I_4\}$	2	$\frac{2}{9} = 22$
$\{I_2, I_5\}$	2	$\frac{2}{9} = 22$
$\{I_3, I_4\}$	0	$\frac{0}{9} = 0 \times$
$\{I_3, I_5\}$	1	$\frac{1}{9} = 11.11 \times$
$\{I_4, I_5\}$	0	$\frac{0}{9} = 0 \times$

Item	frequency	support
$\{I_1, I_2\}$	4	$\frac{4}{9} = 44$
$\{I_1, I_4\}$	4	$\frac{4}{9} = 44$
$\{I_1, I_5\}$	2	$\frac{2}{9} = 22$
$\{I_2, I_3\}$	4	$\frac{4}{9} = 44$
$\{I_2, I_4\}$	2	$\frac{2}{9} = 22$
$\{I_2, I_5\}$	2	$\frac{2}{9} = 22$

Item	frequency	support
$\{I_1, I_2, I_4\}$	2	$\frac{2}{9} \times 100 = 22$
$\{I_1, I_2, I_5\}$	2	$\frac{2}{9} = 22$
$\{I_1, I_2, I_4\}$	1	$\frac{1}{9} = 11.11$ ×
$\{I_1, I_3, I_5\}$	1	$\frac{1}{9} = 11.11$ ×
~~$\{I_1, I_2, I_3\}$~~		
$\{I_2, I_3, I_4\}$	0	$\frac{0}{9} = 0$ ×
$\{I_2, I_3, I_5\}$	1	$\frac{1}{9} = 11.11$ ×
$\{I_2, I_4, I_5\}$	0	$\frac{0}{9} = 0$ ×

Item	frequency	support
$\{I_1, I_2, I_3\}$	2	$\frac{2}{9} = 22$
$\{I_1, I_2, I_5\}$	2	$\frac{2}{9} = 22$

Item	frequency	Support
$\{I_1, I_2, I_3, I_5\}$	1	$\frac{1}{9} = 4.4$ ✗

It does not ~~support~~ satisfy support count. Consider the previous frequency sets

1). $\{I_1, I_2, I_3\}$ — 2/9

2) $\{I_1, I_2, I_5\}$ — 2/9

Confidence

1) $\{I_1, I_2, I_3\}$

Subset → supp-count $I_1 \cup I_2 \cup I_3$ / sup-(I_1)

// non empty sets

R_1 $\{I_1\} \longrightarrow \{I_2, I_3\}$ → $\frac{2/9}{6/9} \times 100 = 33.33$ ✗

R_2 $\{I_2\} \longrightarrow \{I_1, I_3\}$ → $\frac{2/9}{7/9} \times 100 = 28.57$ ✗

R_3 $\{I_3\} \longrightarrow \{I_1, I_2\}$ → $\frac{2/9}{6/9} \times 100 = 33.33$ ✗

R_4 $\{I_1, I_2\} \longrightarrow \{I_3\}$ → $\frac{2/9}{4/9} \times 100 = 50$ ✗

R_5 $\{I_1, I_3\} \longrightarrow \{I_2\}$ → $\frac{2/9}{4/9} \times 100 = 50$ ✗

R_6 $\{I_2, I_3\} \longrightarrow \{I_1\}$ → $\frac{2/9}{4/9} \times 100 = 50$ ✗

R_7 $\{I_1, I_2, I_3\} \longrightarrow \{\phi\}$ → ~~2~~ ϕ

2) $\{I_1, I_2, I_5\}$

// non empty sets

Confidence

R_8 $\{I_1\} \longrightarrow \{I_2, I_5\}$ → $\frac{2/9}{6/9} \times 100 = 33.33$ ✗

R_9 $\{I_2\} \longrightarrow \{I_1, I_5\}$ → $\frac{2/9}{7/9} \times 100 = 28.57$ ✗

R_{10} $\{I_5\} \longrightarrow \{I_1, I_2\}$ → $\frac{2/9}{2/9} \times 100 = 100$ ✓

R_{11} $\{I_1, I_2\} \longrightarrow \{I_5\}$ → $\frac{2/9}{4/9} \times 100 = 50$ ✗

R_{12} $\{I_1, I_5\} \longrightarrow \{I_2\}$ → $\frac{2/9}{2/9} \times 100 = 100$ ✓

R_{13} $\{I_2, I_5\} \rightarrow \{I_1\} \rightarrow \frac{2/9}{2/9} \times 100 = 100$

R_{14} $\{I_1, I_2, I_5\} \rightarrow \emptyset \rightarrow \emptyset$

R_{10}, R_{12}, R_{13} satisfy the minimum confidence threshold so these are strong association rules

Eg:
Generate FP tree for the following transaction Dataset.

minimum support = 30%

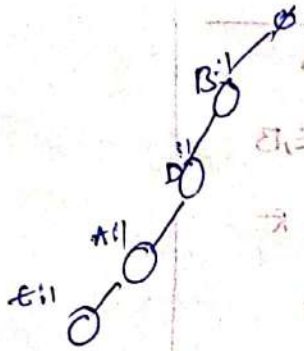
Tr. ID	Items.
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

Item set	frequency	Support count
A	5	5/8 – 62
B	6	6/8 – 75
C	3	3/8 – 37
D	6	6/8 – 75
E	4	4/8 = 50%

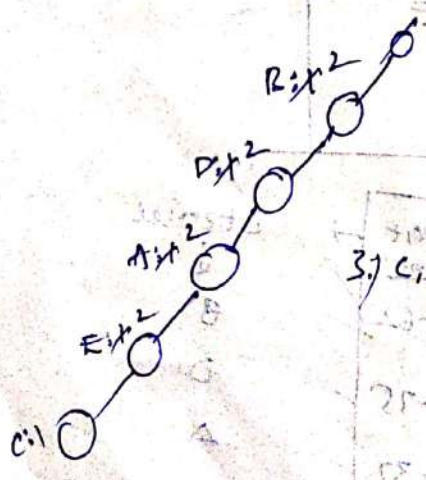
Itemset
↓
B
D
A
E
C

Itemset	frequency	Support count
B	6	$\frac{6}{8} = 75$
D	6	$\frac{6}{8} = 75$
A	5	$\frac{5}{8} = 50$
E	4	$\frac{4}{8} = 62$
C	3	$\frac{3}{8} = 37$

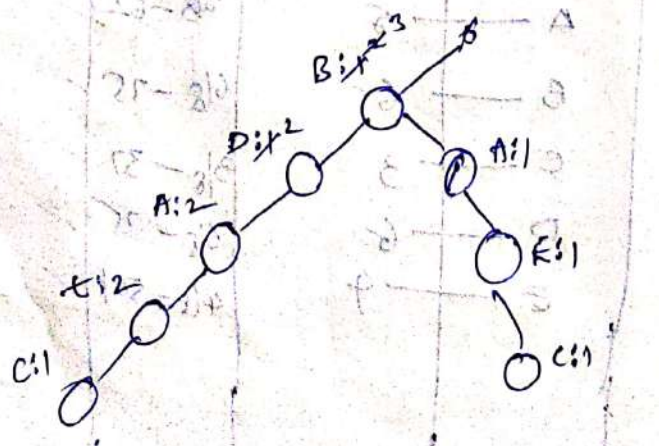
1. E, A, D, B → B D A E



2. D, A, C, E, B → B, D, A, E, C

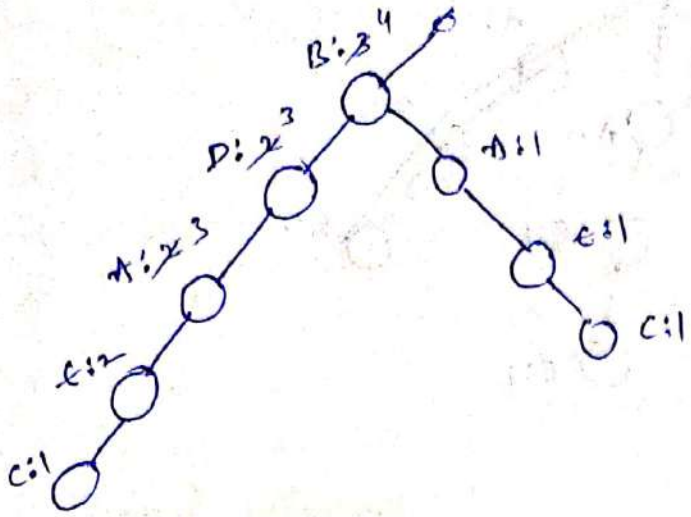


3. C, A, B, E → B, A, E, C

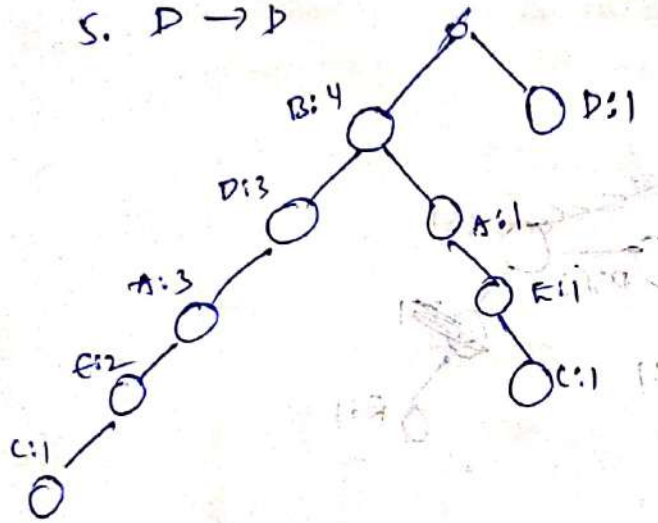


Q.

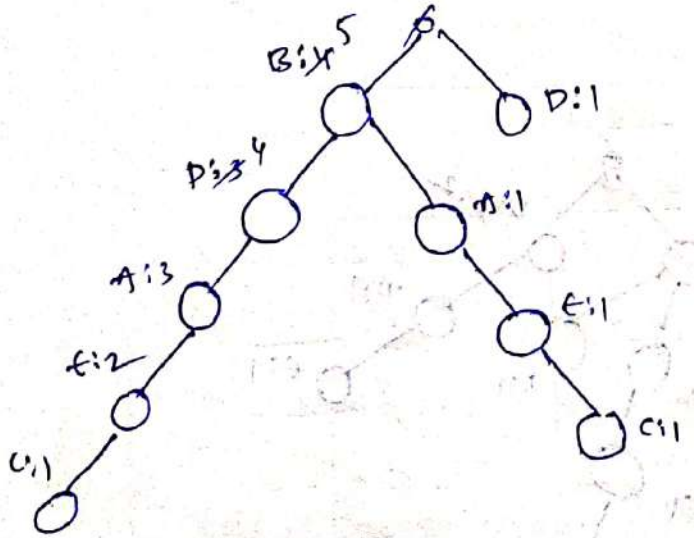
4. B, A, D → B, D, A



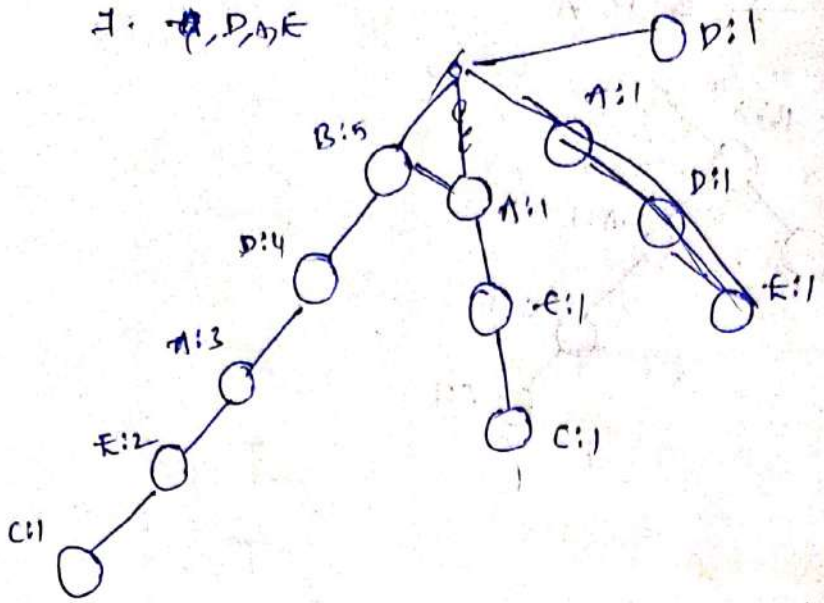
5. D → D



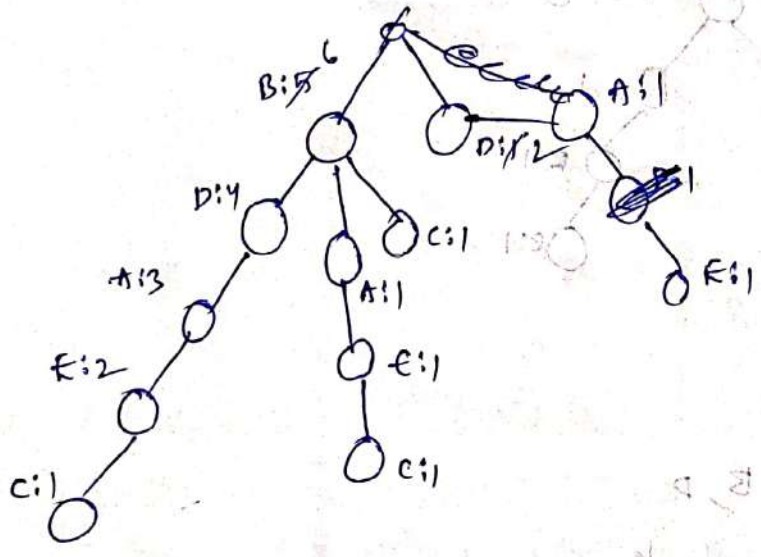
6. D, B → B, D



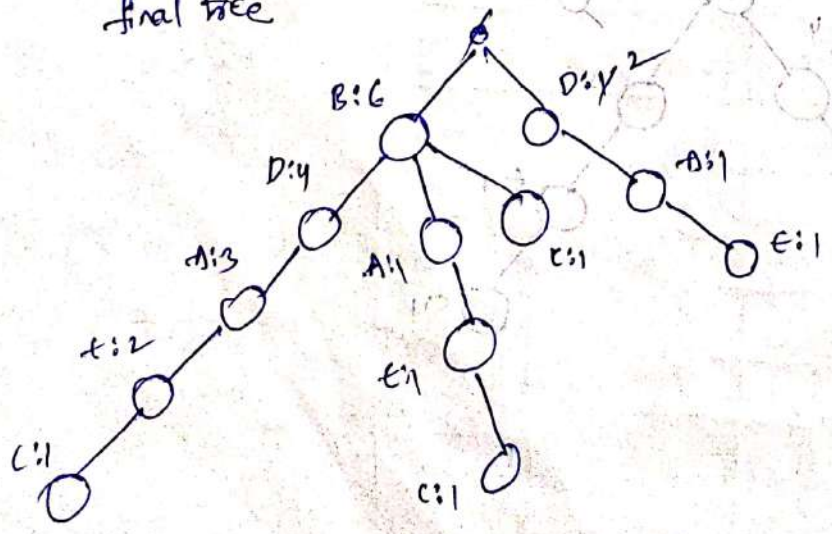
7. A, D, A, E



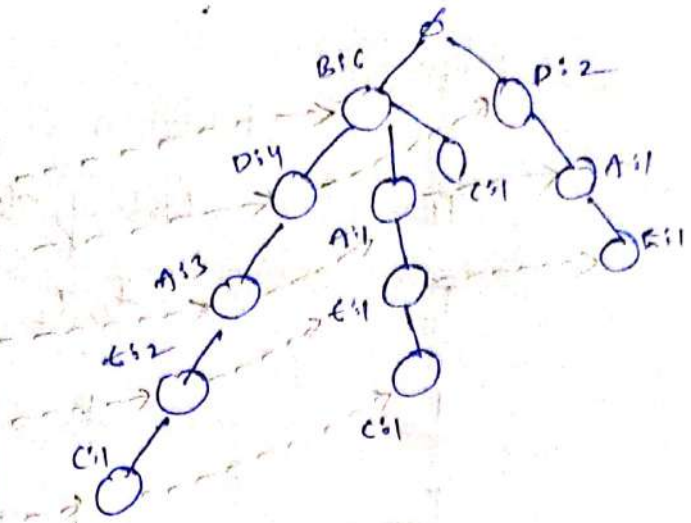
8. B, C → B, C



final tree



Item	support count frequency	node link
B	6	•
D	6	•
A	5	•
E	4	•
C	3	•



Conditional database:-

Item	Conditional pattern base	condition FP tree	Frequent patterns generated.
C	{B, D, A, E:1} {B, A, E:1} {B:1}	{B:3, }	{B, C:3}
E	{D, A:1} {B, D, A:1} {B, A:1}	{A:4, B:3, D:3,}	{A, E:3} {B, E:3} {D, B:3} {A, B

Eclat Algorithm:-

Minimum support = 2

TID	List of Item IDs
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

ECLAT - Equivalent class transformation.

Vertical dataformat :- {item : TID -set}

Horizontal data format : { TID : Itemset}

Vertical data format:

Item set	TID-set
I_1	$\{T_{100}, T_{400}, T_{500}, T_{700}, T_{800}, T_{900}\}$
I_2	$\{T_{100}, T_{200}, T_{300}, T_{400}, T_{600}, T_{800}, T_{900}\}$
I_3	$\{T_{300}, T_{500}, T_{600}, T_{700}, T_{800}, T_{900}\}$
I_4	$\{T_{200}, T_{400},\}$
I_5	$\{T_{100}, T_{800}\}$

Finding 2 item set in vertical data format:

Item set	TID-set
$\{I_1, I_2\}$	$\{T_{100}, T_{400}, T_{800}, T_{900}\}$
$\{I_1, I_3\}$	$\{T_{500}, T_{700}, T_{800}, T_{900}\}$
$\{I_1, I_4\}$	$\{T_{400}\}$ X
$\{I_1, I_5\}$	$\{T_{100}, T_{800}\}$
$\{I_2, I_3\}$	$\{T_{300}, T_{600}, T_{800}, T_{900}\}$
$\{I_2, I_4\}$	$\{T_{200}, T_{400}\}$
$\{I_2, I_5\}$	$\{T_{100}, T_{800}\}$
$\{I_3, I_4\}$	ϕ X
$\{I_3, I_5\}$	$\{T_{800}\}$ X
$\{I_4, I_5\}$	ϕ X

Find 3 item set in vertical data format

Itemset	TID-set
$\{I_1, I_2, I_3\}$	$\{T_{800}, T_{900}\}$
$\{I_1, I_2, I_4\}$	$\{T_{400}\}$ X
$\{I_1, I_3, I_5\}$	$\{T_{800}\}$ X
$\{I_1, I_2, I_5\}$	$\{T_{100}, T_{800}\}$
$\{I_2, I_3, I_4\}$	ϕ X
$\{I_2, I_3, I_5\}$	$\{T_{800}\}$ X
$\{I_2, I_4, I_5\}$	ϕ X

Find 4 item set in vertical data format

Item set	Tid-set
$\{I_1, I_2, I_3, I_5\}$	$\{T800\}$