NAME: MD EZAZ ANWAR ROLL NO.: IIT2018004

COURSE: DATA MINING SEMESTER: VI

ASSIGNMENT - 2

1. Apply apriori algorithm on the following datasets.

A.

TID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

(i) Support = 20%, Confidence = 20%

given, support count = $\frac{20}{100} \times 9 = 1.8 \approx 2$

C_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Scan D for count of each candidate →

Compare candidate support count with minimum support count →

L_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Generate C_2 candidates from L_1 and scan D for count of each candidate →

C_2

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Compare candidate support count with minimum support count →

L_2

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Generate C_3 candidates from L_2 and scan D for count of each candidate →

C_3

Itemset	Sup. Count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate sup. count with min. sup. count →

L_3

Itemset	Sup. Count
{I1, I2, I3}	2
{I1, I2, I5}	2

When we generate C_4 using $L_3 \bowtie L_3$, we get $\{\{I1, I2, I3, I5\}\}$, where itemset $\{I1, I2, I3, I5\}$ is pruned because its subset $\{I2, I3, I5\}$ is not frequent. Thus, $C_4 = \Phi$ and algorithm terminates.

Now, we have frequent itemsets of size greater than 2,

$L = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}, \{I1, I2, I3\}, \{I1, I2, I5\}\}$

so, the association rules we get are,

I. $\{I1\} \Rightarrow \{I2\}$ confidence = 4/6 = 67%.

II. $\{I2\} \Rightarrow \{I1\}$ confidence = 4/7 = 57%.

III. $\{I1\} \Rightarrow \{I3\}$ confidence = 4/6 = 67%.

IV. $\{I3\} \Rightarrow \{I1\}$ confidence = 4/6 = 67%.

V. $\{I1\} \Rightarrow \{I5\}$ confidence = 2/6 = 33%.

VI. $\{I5\} \Rightarrow \{I1\}$ confidence = 2/2 = 100%.

VII. $\{I2\} \Rightarrow \{I3\}$ confidence = 4/7 = 57%.

VIII. $\{I3\} \Rightarrow \{I2\}$ confidence = 4/6 = 67%.

IX. $\{I2\} \Rightarrow \{I4\}$ confidence = 2/7 = 29%.

X. $\{I4\} \Rightarrow \{I2\}$ confidence = 2/2 = 100%.

XI. $\{I2\} \Rightarrow \{I5\}$ confidence = 2/7 = 29%.

XII. $\{I5\} \Rightarrow \{I2\}$ confidence = 2/2 = 100%.

XIII. $\{I1, I2\} \Rightarrow \{I3\}$ confidence = 2/4 = 50%.

XIV. $\{I1, I3\} \Rightarrow \{I2\}$ confidence = 2/4 = 50%.

XV. $\{I2, I3\} \Rightarrow \{I1\}$ confidence = 2/4 = 50%.

XVI. $\{I3\} \Rightarrow \{I1, I2\}$ confidence = 2/6 = 33%.

XVII. $\{I2\} \Rightarrow \{I1, I3\}$ confidence = 2/7 = 29%.

XVIII. $\{I1\} \Rightarrow \{I2, I3\}$ confidence = 2/6 = 33%.

XIX. $\{I1, I2\} \Rightarrow \{I5\}$ confidence = 2/4 = 50%.

XX. $\{I1, I5\} \Rightarrow \{I2\}$ confidence = 2/2 = 100%.

XXI. $\{I2, I5\} \Rightarrow \{I1\}$ confidence = 2/2 = 100%.

XXII. $\{I5\} \Rightarrow \{I1, I2\}$ confidence = 2/2 = 100%.

XXIII. $\{I2\} \Rightarrow \{I1, I5\}$ confidence = 2/7 = 29%.

XXIV. $\{I1\} \Rightarrow \{I2, I5\}$ confidence = 2/6 = 33%.

All the association rules have confidence ≥ minimum confidence. Hence, all the rules are strong association rules.

(ii) Support = 30%, Confidence = 30%.

Given, support count = $\frac{30}{100} \times 9 = 2.7 \approx 3$

(a) Scan D for count of each candidate,

C_1

Itemset	{I1}	{I2}	{I3}	{I4}	{I5}
Sup. Count	6	7	6	2	2

Compare Support count with minimum support count →

L_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6

Generate C_2 from L_1 and scan D for count →

C_2

Itemset	Count
{I1, I2}	4
{I1, I3}	4
{I2, I3}	4

Compare cand. sup. count with min. sup. count →

L_2

Itemset	Count
{I1, I2}	4
{I1, I3}	4
{I2, I3}	4

Generate C_3 from L_2 and scan D for count →

C_3

Itemset	Sup. Count
{I1, I2, I3}	2

When we generate L_3 from C_3 we get $L_3 = \phi$ and algorithm terminates.

So, we have, $L = \{\{I1, I2\}, \{I1, I3\}, \{I2, I3\}\}$

Hence, association rules,

I. $\{I1\} \Rightarrow \{I2\}$ confidence = 4/6 = 67%.

II. $\{I2\} \Rightarrow \{I1\}$ confidence = 4/7 = 57%.

III. $\{I1\} \Rightarrow \{I3\}$ confidence = 4/6 = 67%.

IV. $\{I3\} \Rightarrow \{I1\}$ confidence = 4/6 = 67%.

V. $\{I2\} \Rightarrow \{I3\}$ confidence = 4/7 = 57%.

VI. $\{I3\} \Rightarrow \{I2\}$ confidence = 4/6 = 67%.

All the association rules have confidence ≥ 30%. Hence, all rules are strong association rules.

B.

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

(i) Support = 40%. , Confidence = 40%.

We have, support count = $\frac{40}{100} \times 5 = 2$

STEP I : K = 1

Scan D for count of each candidate.

C_1

Itemset	{E}	{K}	{M}	{N}	{O}	{Y}	{D}	{A}	{C}	{U}	{I}
Count	4	5	3	2	3	3	1	1	2	1	1

Compare candidate support count with minimum support count

L_1

Itemset	{E}	{K}	{M}	{N}	{O}	{Y}	{C}
Count	4	5	3	2	3	3	2

STEP II : K = 2

Generate C_2 using L_1 and count each candidate by scanning D.

C_2

Itemset	{E,K}	{E,M}	{E,N}	{E,O}	{E,Y}	{E,C}	{K,M}	{K,N}	{K,O}	{K,Y}	{K,C}
Count	4	2	2	3	2	1	2	2	3	3	2

C_2

Itemset	{M,N}	{M,O}	{M,Y}	{M,C}	{N,O}	{N,Y}	{N,C}	{O,Y}	{O,C}	{Y,C}
Count	1	1	2	1	2	2	0	2	1	1

Compare candidate support count with minimum support count.

L_2

Itemset	{E,K}	{E,M}	{E,N}	{E,O}	{E,Y}	{K,M}	{K,N}
Count	4	2	2	3	2	2	2

L_2

Itemset	{K,O}	{K,Y}	{K,C}	{M,Y}	{N,O}	{N,Y}	{O,Y}
Count	3	3	2	2	2	2	2

STEP - III : K = 3

Generate C_3 using L_2 and scan D for count of each candidate.

C_3

Itemset	{E,K,M}	{E,K,N}	{E,K,O}	{E,K,Y}	{E,M,Y}	{E,N,O}	{E,N,Y}
Count	2	2	3	2	1	2	2

C_3

Itemset	{E,O,Y}	{K,M,Y}	{K,N,O}	{K,N,Y}	{K,O,Y}	{N,O,Y}
Count	2	2	2	2	2	2

Compare candidate support count with minimum support count.

L_3

Itemset	{E,K,M}	{E,K,N}	{E,K,O}	{E,K,Y}	{E,N,O}	{E,N,Y}
Count	2	2	3	2	2	2

L_3

Itemset	{E,O,Y}	{K,M,Y}	{K,N,O}	{K,N,Y}	{K,O,Y}	{N,O,Y}
Count	2	2	2	2	2	2

STEP - IV : K = 4

Generate C_4 using L_3 and scan D for count of each candidate.

C_4

Itemset	{E,N,O,Y}	{E,K,N,O}	{E,K,O,Y}	{K,N,O,Y}	{E,K,N,Y}
Count	2	2	2	2	2

Compare candidate support count with minimum support count.

L_4

Itemset	{E,N,O,Y}	{E,K,N,O}	{E,K,O,Y}	{K,N,O,Y}	{E,K,N,Y}
Count	2	2	2	2	2

STEP-V : K=5

Generate C_5 using L_4 and scan D for count of each candidate.

C_5

Itemset	Count
{E,K,N,O,Y}	2

Compare candidate support count with minimum sup. count.

L_5

Itemset	Count
{E,K,N,O,Y}	2

When we generate $C_6 = L_5 \bowtie L_5$ we get $C_6 = \phi$, hence algorithm terminates.

So, we have frequent itemset,

$L = L_2 \cup L_3 \cup L_4 \cup L_5$

There are 14 elements in L_2, so it will give $14 \times {}^2C_1 = 28$ rules

Similarly L_3 will give $12 \times {}^3C_1 \times 2 = 72$ rules

And, L_4 will give $5 \times ({}^4C_1 \times 2 + {}^4C_2) = 70$ rules

And, L_5 will give $1 \times ({}^5C_1 \times 2 + {}^5C_2) = 30$ rules

i.e. a total of 200 association rules. Hence, only some rules are shown below.

Using, {E,K} we get two rules

I. {E} ⇒ {K} confidence = 4/4 = 100%.

II. {K} ⇒ {E} confidence = 4/5 = 80%.

Using, {E,K,O} we get six rules

III. {E,K} ⇒ {O} confidence = 3/4 = 75%.

IV. {E,O} ⇒ {K} confidence = 3/3 = 100%.

V. {K,O} ⇒ {E} confidence = 3/3 = 100%.

VI. {O} ⇒ {E,K} confidence = 3/3 = 100%.

VII {K} ⇒ {E,O} confidence = 3/5 = 60%.

VIII. {E} ⇒ {K,O} confidence = 3/4 = 75%.

Using {E,K,O,Y}, we will have 14 rules, some of which are,

IX. {E,K,O} ⇒ {Y} confidence = 2/3 = 67%.

X. {K} ⇒ {E,O,Y} confidence = 2/5 = 40%.

XI. {E,O} ⇒ {K,Y} confidence = 2/3 = 67%.

Using {E,K,N,O,Y}, we will have 30 rules, some of which are,

XII. {E,K,N,O} ⇒ {Y} confidence = 2/2 = 100%.

XIII. {E,K,N} ⇒ {O,Y} confidence = 2/2 = 100%.

XIV. {E,K} ⇒ {N,O,Y} confidence = 2/4 = 50%.

XV. {E} ⇒ {K,N,O,Y} confidence = 2/4 = 50%.

All the rules have confidence ≥ min_confidence. So, all rules are strong association rules.

(ii). Support = 3 Confidence = 90%.

Given, support = 3

STEP-I : K = 1

Scan D to find the count of each candidate.

C_1

Itemset	{E}	{K}	{M}	{N}	{O}	{Y}	{D}	{A}	{C}	{U}	{I}
Count	4	5	3	2	3	3	1	1	2	1	1

Compare candidate count with minimum support count.

L_1

Itemset	{E}	{K}	{M}	{O}	{Y}
Count	4	5	3	3	3

STEP-II : K = 2

Scan D to find the count of each candidate generated in C_2.

C_2

Itemset	{E,K}	{E,M}	{E,O}	{E,Y}	{K,M}	{K,O}	{K,Y}	{M,O}	{M,Y}	{O,Y}
Count	4	2	3	2	3	3	3	1	2	2

Compare candidate count with minimum support count.

L_2

Itemset	{E,K}	{E,O}	{K,M}	{K,O}	{K,Y}
Count	4	3	3	3	3

STEP-III : K=3

Generate C_3 using L_2 and scan D to find the count of each cand.

C_3

Itemset	Count
{E,K,O}	3

Compare candidate support count with min. support count.

L_3

Itemset	Count
{E,K,O}	3

When we generate C_4 using L_3 we get $C_4 = \phi$ and algorithm terminates.

Now, we have,

$L = \{\{E,K\}, \{E,O\}, \{K,M\}, \{K,O\}, \{K,Y\}, \{E,K,O\}\}$

So, the association rules are,

i. $\{E\} \Rightarrow \{K\}$ confidence = 4/4 = 100%.

ii. $\{K\} \Rightarrow \{E\}$ confidence = 4/5 = 80%.

iii. $\{E\} \Rightarrow \{O\}$ confidence = 3/4 = 75%.

iv. $\{O\} \Rightarrow \{E\}$ confidence = 3/3 = 100%.

v. $\{K\} \Rightarrow \{M\}$ confidence = 3/5 = 60%.

vi. $\{M\} \Rightarrow \{K\}$ confidence = 3/3 = 100%.

vii. $\{K\} \Rightarrow \{O\}$ confidence = 3/5 = 60%.

viii. $\{O\} \Rightarrow \{K\}$ confidence = 3/3 = 100%.

ix. $\{K\} \Rightarrow \{Y\}$ confidence = 3/5 = 60%.

x. $\{Y\} \Rightarrow \{K\}$ confidence = 3/3 = 100%.

xi. $\{E,K\} \Rightarrow \{O\}$ confidence = 3/4 = 75%.

xii. $\{O\} \Rightarrow \{E,K\}$ confidence = 3/3 = 100%.

xiii. $\{E,O\} \Rightarrow \{K\}$ confidence = 3/3 = 100%.

xiv. $\{K\} \Rightarrow \{E,O\}$ confidence = 3/5 = 60%.

xv. $\{O,K\} \Rightarrow \{E\}$ confidence = 3/3 = 100%.

xvi. $\{E\} \Rightarrow \{O,K\}$ confidence = 3/4 = 75%.

All the above rules are strong association rules, since every rule has conf $\geq$ 20%.

2. Apply FP-growth algorithm on the following datasets.

A.

Transaction ID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

Support = 20% and Confidence = 20%

Given, support count = $0.2 \times 9 = 1.8 \approx 2$

scan the database to find the count of each item.

Item	I1	I2	I3	I4	I5
count	6	7	6	2	2

The set L is constructed in the order of decreasing support count

$L = \{\{I2:7\}, \{I1:6\}, \{I3:6\}, \{I4:2\}, \{I5:2\}\}$

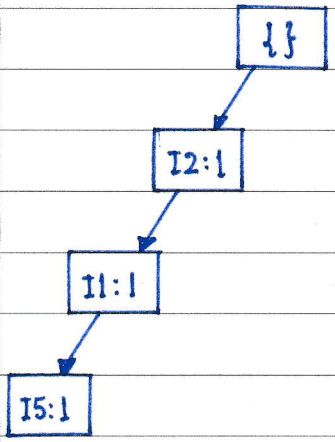
Now, ordered itemset is built by scanning database and L.

Transaction ID	Ordered Itemset
T1	{I2, I1, I5}
T2	{I2, I4}
T3	{I2, I3}
T4	{I2, I1, I4}
T5	{I1, I3}

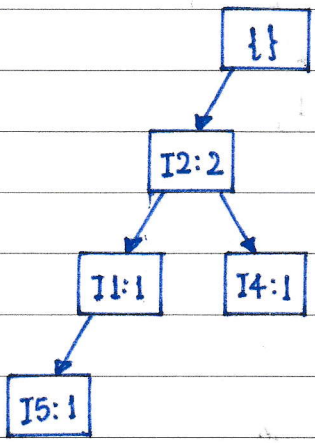
Transaction ID	Ordered Itemset
T6	{I2, I3}
T7	{I1, I3}
T8	{I1, I1, I3, I5}
T9	{I2, I1, I3}

Now, we will make the FP-tree.

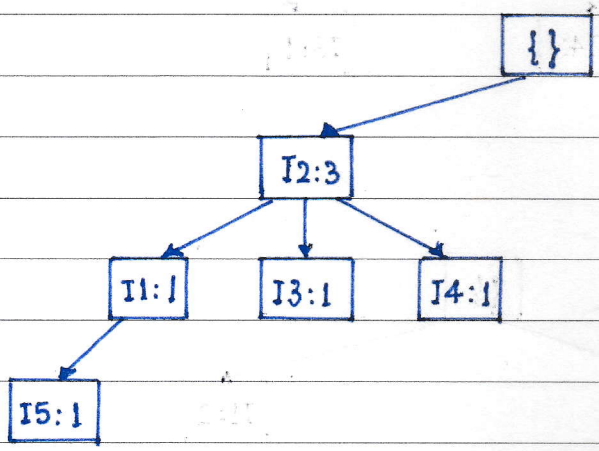
(i) Inserting {I2, I1, I5},



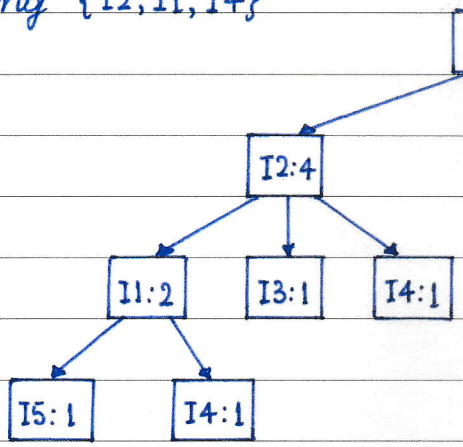
(ii) Inserting {I2, I4}



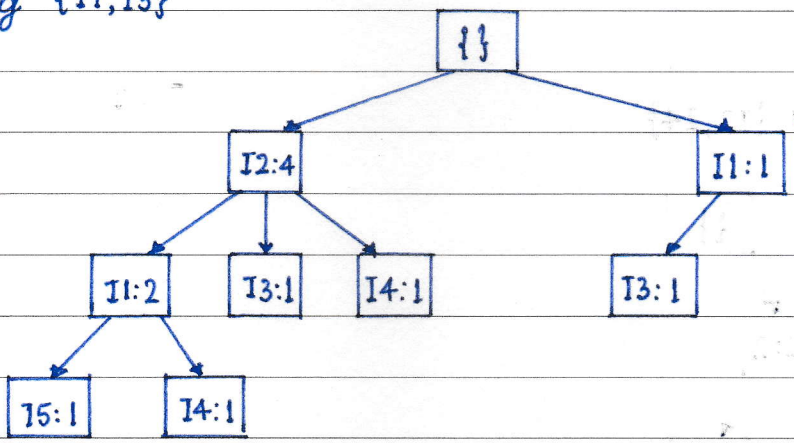
(iii) Inserting {I2, I3}



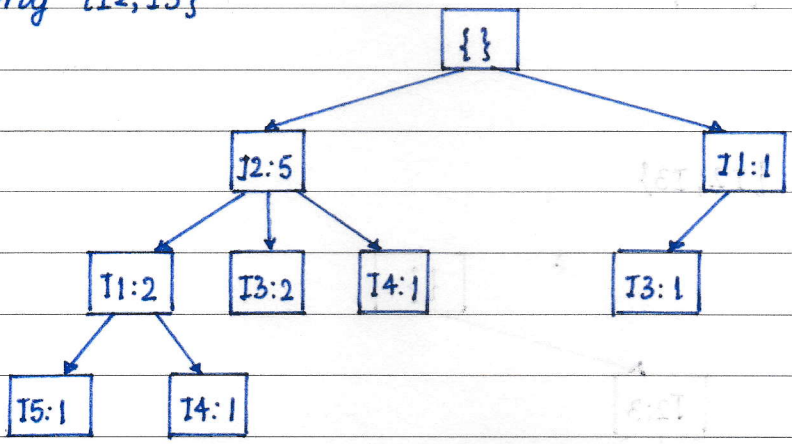
(iv) Inserting {I2, I1, I4}



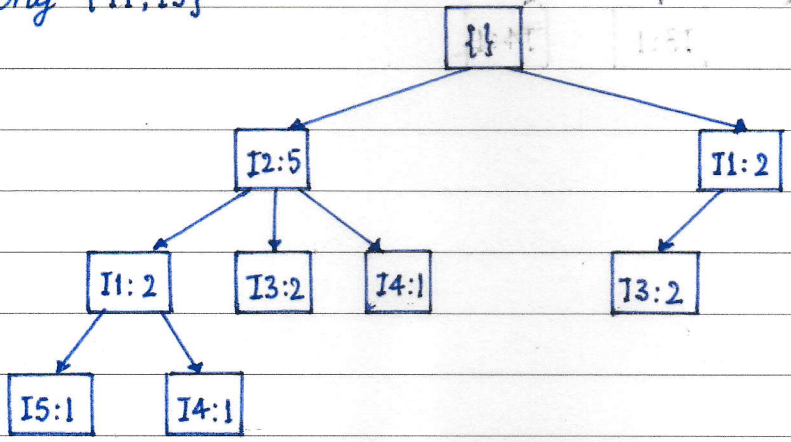
(v) Inserting {I1, I3}



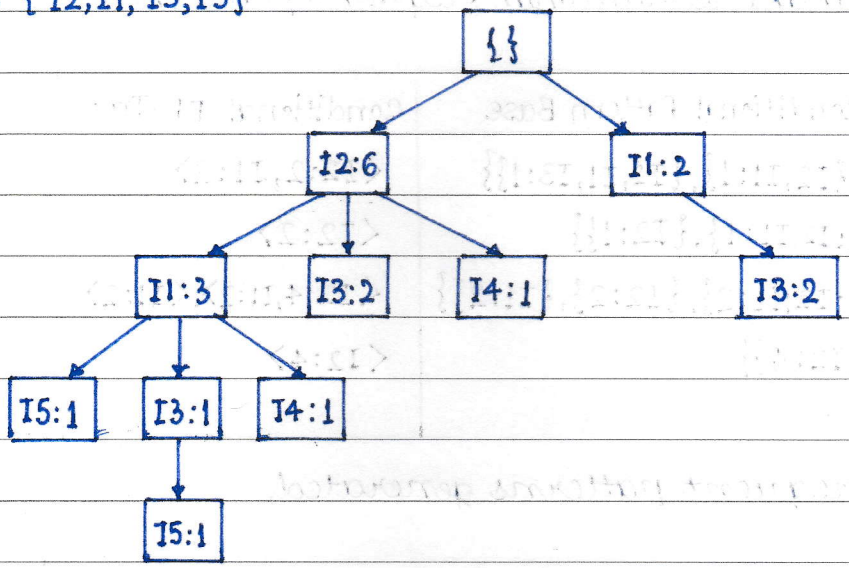
(vi) Inserting {I2, I3}



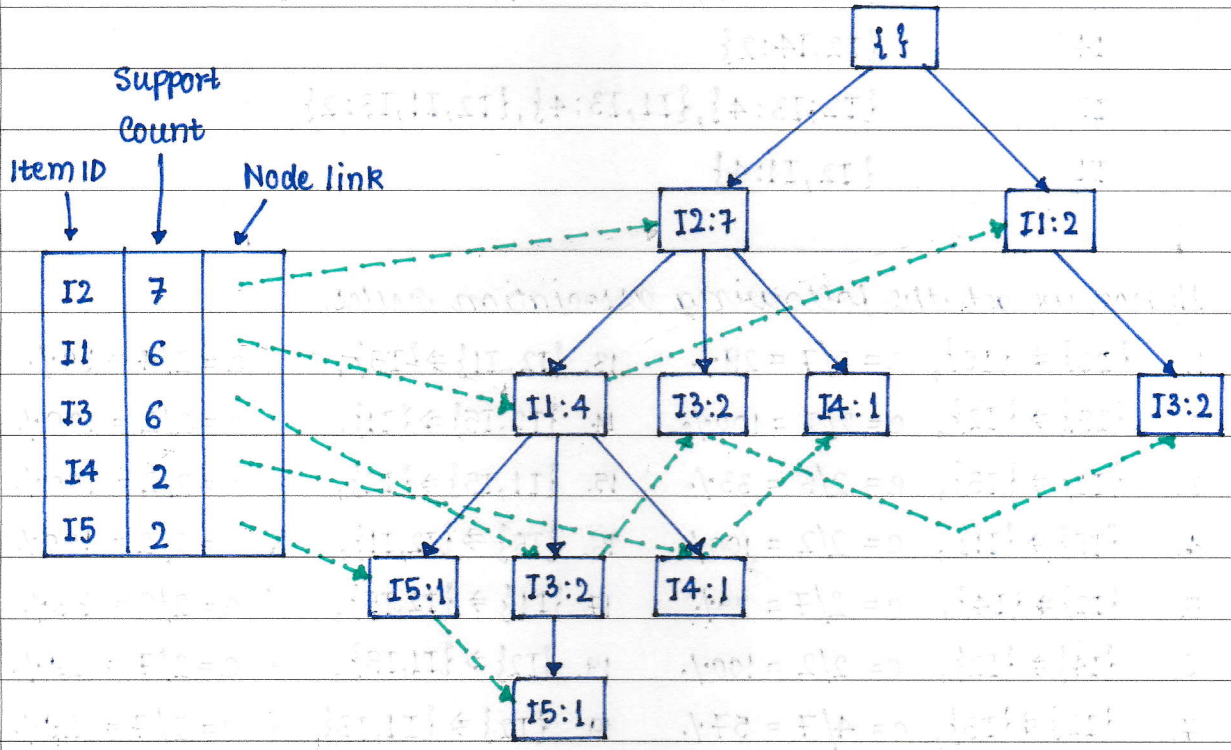
(vii) Inserting {I1, I3}



(viii) Insert {I2, I1, I3, I5}



(ix) After inserting {I2, I1, I3}, the final FP-tree,



Now, we generate Conditional Pattern Base,

Items	Conditional Pattern Base
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}
I4	{{I2, I1: 1}, {I2: 1}}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}
I1	{{I2: 4}}

Now, we make conditional frequent pattern tree,

Items	Conditional Pattern Base	Conditional FP-Tree
I5	{{I2,I1:1},{I2,I1,I3:1}}	<I2:2, I1:2>
I4	{{I2,I1:1},{I2:1}}	<I2:2>
I3	{{I2;I1:2},{I2:2},{I1:2}}	<I2:4,I1:2>,<I1:2>
I1	{{I2:4}}	<I2:4>

Hence, frequent patterns generated,

Items	Frequent Pattern Generated
I5	{I2,I5:2},{I2,I5:2},{I2,I1,I5:2}
I4	{I2,I4:2}
I3	{I2;I3:4},{I1,I3:4},{I2,I1,I3:2}
I1	{I2,I1:4}

Hence, we get the following association rules,

1. {I2} ⇒ {I5}, c = 2/7 = 29%.
2. {I5} ⇒ {I2}, c = 2/2 = 100%.
3. {I1} ⇒ {I5}, c = 2/6 = 33%.
4. {I5} ⇒ {I1}, c = 2/2 = 100%.
5. {I2} ⇒ {I4}, c = 2/7 = 29%.
6. {I4} ⇒ {I2}, c = 2/2 = 100%.
7. {I2} ⇒ {I3}, c = 4/7 = 57%.
8. {I3} ⇒ {I2}, c = 4/6 = 67%.
9. {I1} ⇒ {I3}, c = 4/6 = 67%.
10. {I3} ⇒ {I1}, c = 4/6 = 67%.
11. {I2} ⇒ {I1}, c = 4/7 = 57%.
12. {I1} ⇒ {I2}, c = 4/6 = 67%.
13. {I2, I1} ⇒ {I5}, c = 2/4 = 50%.
14. {I2,I5} ⇒ {I1}, c = 2/2 = 100%.
15. {I1,I5} ⇒ {I2}, c = 2/2 = 100%.
16. {I5} ⇒ {I2,I1}, c = 2/2 = 100%.
17. {I1} ⇒ {I2,I5}, c = 2/6 = 33%.
18. {I2} ⇒ {I1,I5}, c = 2/7 = 29%.
19. {I2} ⇒ {I1,I3}, c = 2/7 = 29%.
20. {I1} ⇒ {I2,I3}, c = 2/6 = 33%.
21. {I3} ⇒ {I1,I2}, c = 2/6 = 33%.
22. {I1,I2} ⇒ {I3}, c = 2/4 = 50%.
23. {I1,I3} ⇒ {I2}, c = 2/4 = 50%.
24. {I2,I3} ⇒ {I1}, c = 2/4 = 50%.

All the association rules have confidence ≥ 20%. So, all rules are strong association rules.

B.

Transaction ID	Items
T1	E, K, M, N, O, Y
T2	D, E, K, N, O, Y
T3	A, E, K, M
T4	C, K, M, U, Y
T5	C, E, I, K, O, O

Support = 3 and confidence = 20%

Scan the database to find the count of each item.

Item	E	K	M	N	O	Y	C	D	A	U	I
Count	4	5	3	2	3	3	2	1	1	1	1

The set L is constructed in the order of decreasing support count.

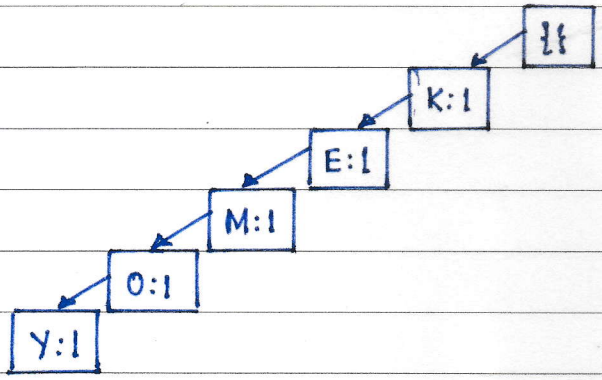
L = {{K:5}, {E:4}, {M:3}, {O:3}, {Y:3}}

Now, the ordered itemset is built by scanning the database,

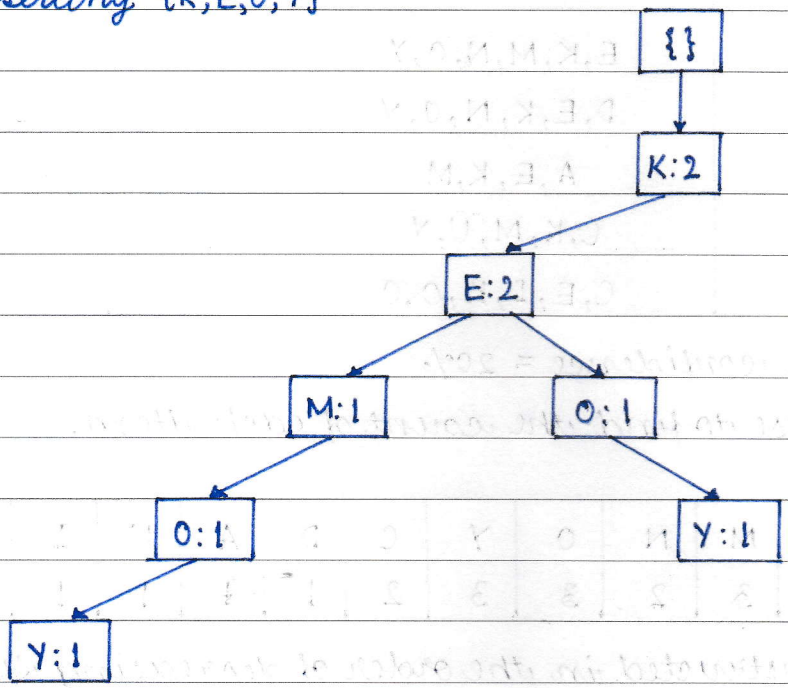
Transaction ID	Ordered Itemset
T1	{K, E, M, O, Y}
T2	{K, E, O, Y}
T3	{K, E, M}
T4	{K, M, Y}
T5	{K, E, O}

Now, we will make the FP-Tree

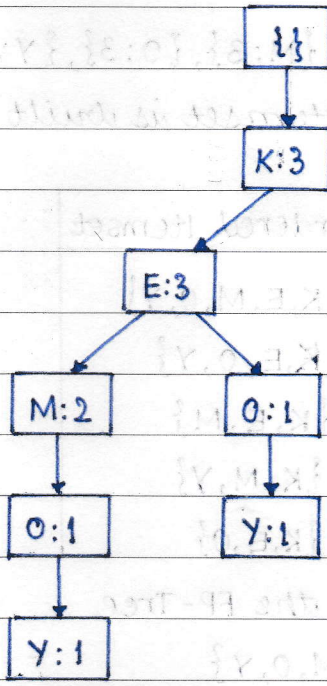
(i) Inserting {K, E, M, O, Y}



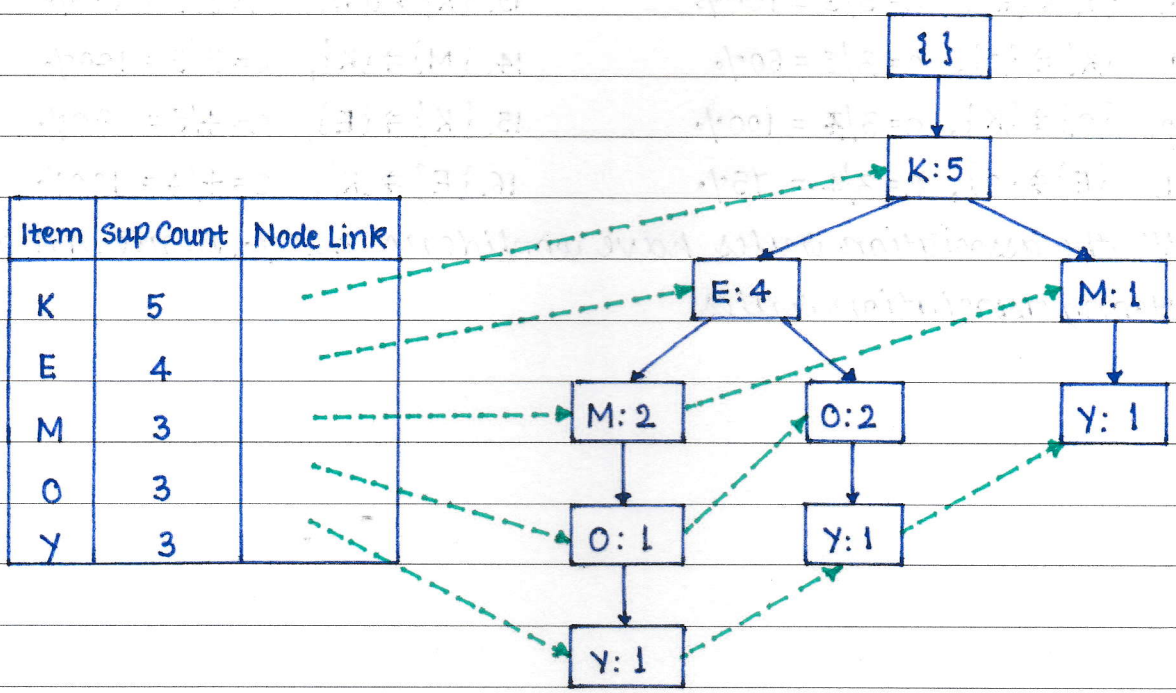
(ii) Inserting {K, E, O, Y}



(iii) Inserting {K, E, M}



(iv) Inserting {K, M, Y} and {K, E, O}, we have the final tree,



Item	Sup Count	Node Link
K	5	
E	4	
M	3	
O	3	
Y	3	

Therefore, conditional pattern base and conditional FP-tree are,

Item	Conditional Pattern Base	Conditional FP-Tree
Y	{{K,E,M,O:1},{K,E,O:1},{K,M:1}}	<K:3>
O	{{K,E,M:1},{K,E:2}}	<K:3, E:3>
M	{{K,E:2},{K:1}}	<K:3>
E	{{K:4}}	<K:4>

So, frequent patterns generated,

Item	Frequent Patterns
Y	{K, Y:3}
O	{K, E, O:3}, {K, O:3}, {E, O:3}
M	{K, M:3}
E	{K, E:4}

Therefore, we have following association rules,

1. {K, E} ⇒ {O}, c = 3/4 = ~~100~~ 75%
2. {K, O} ⇒ {E}, c = 3/3 = 100%
3. {E, O} ⇒ {K}, c = 3/3 = 100%
4. {O} ⇒ {E, K}, c = 3/3 = 100%
5. {E} ⇒ {O, K}, c = 3/4 = 75%
6. {K} ⇒ {O, E}, c = 3/5 = 60%

7. $\{K\} \Rightarrow \{Y\}$, $c = 3/5 = 60\%$
8. $\{Y\} \Rightarrow \{K\}$, $c = 3/3 = 100\%$
9. $\{K\} \Rightarrow \{O\}$, $c = 3/5 = 60\%$
10. $\{O\} \Rightarrow \{K\}$, $c = 3/3 = 100\%$
11. $\{E\} \Rightarrow \{O\}$, $c = 3/4 = 75\%$
12. $\{O\} \Rightarrow \{E\}$, $c = 3/3 = 100\%$
13. $\{K\} \Rightarrow \{M\}$, $c = 3/5 = 60\%$
14. $\{M\} \Rightarrow \{K\}$, $c = 3/3 = 100\%$
15. $\{K\} \Rightarrow \{E\}$, $c = 4/5 = 80\%$
16. $\{E\} \Rightarrow \{K\}$, $c = 4/4 = 100\%$

All the association rules have confidence $\geq 20\%$. Hence, all are strong association rules.

3. Why we use union in the formula of confidence instead of intersection?

The confidence of an association rule $X \Rightarrow Y$ is given by,

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

The point to note here is that X and Y are itemsets not transactions. So, $\text{supp}(X \cup Y)$ is the support where X and Y both itemsets occur. The union is used because here we are defining support in terms of sets of items and not events. We can use intersection if we define events E_X and E_Y, where E_X denotes the event that the transaction contains itemset X. So, we can write, $\text{support}(X \cup Y) = P(E_X \cap E_Y)$.

4. What is the physical significance of confidence and lift?

Confidence of an association rule $X \Rightarrow Y$ is given by,

$$\text{conf}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

Confidence, in simple words, is measure of the possibility the if an item X is bought, then the item Y will also be bought. It is conditional probability of, X and Y both being purchased, given, X is already purchased.

Lift of an association rule $X \Rightarrow Y$ is given by,

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{sup}(Y)} = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)\text{sup}(Y)}$$

Lift is a simple correlation measure, i.e. it siginifies the dependency/independency of the two events, i.e. sale of X and sale of Y. In other words, it is a measure of how sale of X "lifts" the sale of Y. If lift $(X \Rightarrow Y) < 1$, it means the sale of X is going to decrease the sale of Y. If the value of lift is 1, it means, sale of Y is independent of sale of X. If value of lift > 1, it means, sale of X is likely going to increase the sale of Y.

5. What is the role of partitioning and sampling in apriori algorithm?

The partitioning and sampling algorithms are used to increase the efficiency of apriori algorithm. Basic apriori algorithm is slow as it requires complete database scan at each level of frequent itemset generation.

In partioning the database is divided into n-non-overlapping partitions. The frequent itemsets local to each partition is found. Then all the frequent itemsets local to each partion is combined to form candidate itemset. Then the database is scanned for second time to find global frequent itemsets among those candidates.

In sampling, the random sample S of database is picked and frequent itemsets are searched in S instead of complete database. But, since a random sample is searched, it is possible that some frequent itemsets are missed. In this way, we trade off some degree of accuracy against efficiency.