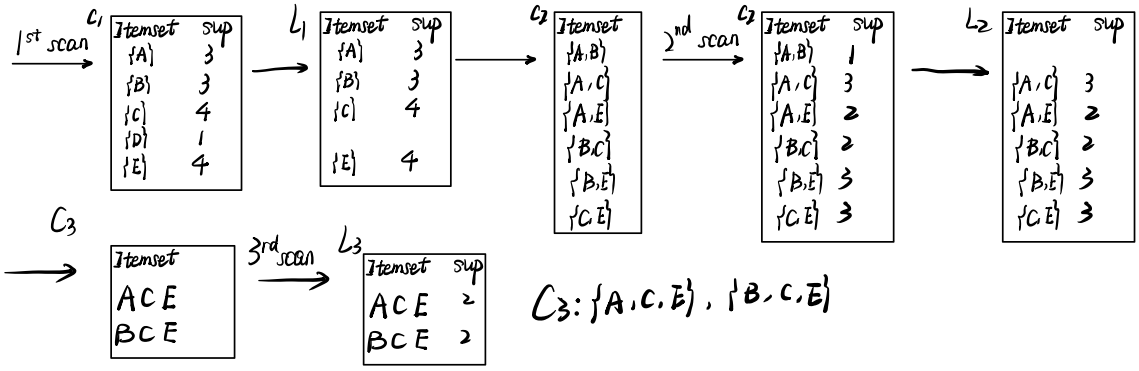
Q1 [15 Marks]

Given the transaction database below, set the minimum support count to 2 and the minimum confidence level to 60% to find the strong association rule. Generate the set C_3 of the candidate 3-itemset , using prunning on Apriori principle.

TID	Item
T1	A,C,D
T2	B,C,E
T3	A,B,C,E
T4	B,E
T5	A,C,E

$sup_{min}=2$



step1: Generate frequent 1-itemsets: {A} {B} {C} {E}

step2: Generate frequent 2-itemsets: {A,C}, {A,E}, {B,C}, {B,E}, {C,E}

step3: Generate frequent 3-itemsets (apply pruning based on the Apriori principle)

C_3: {A,C,E}, {B,C,E}

step4: calculate.

$A\to C$, $A\to E$, $C\to E$, $C\to A$, $E\to A$, $E\to C$, $B\to C$, $B\to E$, $C\to B$, $E\to B$

$\frac{2}{3}$ $\frac{2}{3}$ $\frac{3}{4}$ $\frac{3}{4}$ $\frac{2}{4}$ $\frac{3}{4}$ $\frac{2}{3}$ $\frac{3}{3}$ $\frac{2}{4}$ $\frac{3}{4}$

$A\to C,E$, $C\to A,E$, $E\to A,C$, $C,E\to A$, $A,E\to C$, $A,C\to E$

$\frac{2}{3}$ $\frac{2}{4}$ $\frac{2}{4}$ $\frac{2}{3}$ $\frac{2}{2}$ $\frac{2}{3}$

$B\to C,E$, $C\to B,E$, $E\to B,C$, $C,E\to B$, $B,E\to C$, $B,C\to E$.

$\frac{2}{3}$ $\frac{2}{4}$ $\frac{2}{4}$ $\frac{2}{3}$ $\frac{2}{3}$ $\frac{2}{2}$

confidence > 60%, the strong association rules are:

$A\to C$, $A\to E$, $C\to E$, $C\to A$, $E\to C$, $B\to C$, $B\to E$, $E\to B$

$A\to C,E$, $C,E\to A$, $A,E\to C$ $A,C\to E$, $B\to C,E$, $C,E\to B$ $B,E\to C$, $B,C\to E$.

Q2 [15 Marks]

Reducing the transactions using dynamic hashing and pruning(DHP) algorithm. Set the minimum support count to 2.

Hash function bucket #= h({x y}) = ((order of x)*10+(order of y)) % 7

TID	Item
T1	A,B,C
T2	B,D,E
T3	A,B,D,E
T4	B,E

Items = A, B, C, D, E
Order = 1, 2, 3, 4, 5

The minimum support is 2

Itemset	support
A	2
B	4
C	1
D	2
E	3

C_1

Itemset	support
A	2
B	4
D	2
E	3

L_1

Find all 2-itemset of each transaction

TID	2-itemset
T_1	{AB} {AC} {BC}
T_2	{BD} {BE} {DE}
T_3	{AB} {AD} {AE} {BD} {BE} {DE}
T_4	{BE}

Hash function h({x y}) = ((order of x)·10+(order of y)) %7

Hash table : map the 2-item of transaction into the hash table

bucket 0	1	2	3	4	5	6
{AD}	{AE}	{BC}	{BD} {DE} {BD} {DE}	{BE} {BE} {BE}	{AB} {AB}	{AC}
1	1	1	4	3	2	1

$L_1 \times L_1$	# in the bucket
{A B}	2
{A D}	1
{A E}	1
{B D}	4
{B E}	3
{D E}	4

C_2
{A B}
{B D}
{B E}
{D E}

TID	Item
T1	A,B,C
T2	B,D,E
T3	A,B,D,E
T4	B,E

{A B} → discard
{B D} {B E} {D E} → keep {B D E}
{A B} {B D} {B E} {D E} → keep {B D E}
{B E} → discard

C_2	count	L_2
{A B}	2	{AB}
{B D}	2	{BD}
{B E}	2	{BE}
{D E}	2	{DE}

$D_3 = \{<T_2, B\ D\ E>, <T_3, B\ D\ E>\}$

Q3 [35 Marks]

An itemset X is said to be a frequent itemset if the frequency count of X is at least a given support threshold.

An itemset Y is a proper super-itemset of X if X $\subset$Y and X $\neq$Y.

An itemset X is said to be a closed frequent itemset if (1) X is frequent and (2) there exists no proper super itemset Y of X such that Y is frequent and Y has the same frequency count as X.

An itemset X is said to be a maximal frequent itemset if (1) X is frequent and (2) there exists no proper super-itemset Y of X such that Y is frequent.

Let F be the set of (traditional) frequent itemsets without specifying the frequency of itemsets.

Let F_c be the set of (traditional) frequent itemsets each of which is associated with a frequency in the dataset.

For example, if there are three frequent itemsets, $\{I_1\}$ with frequency 4, $\{I_2\}$ with frequency 5, and $\{I_1, I_2\}$ with frequency 3, F = $\{\{I_1\}, \{I_2\}, \{I_1,I_{12}\}\}$ and Fc = $\{<\{I_1\}, 4>, <\{II_2\}, 5>, <\{I_1, I_2\}, 3>\}$.

Similarly, let C be the set of closed frequent itemsets without specifying the frequency of itemsets.

Let C_c be the set of closed frequent itemsets each of which is associated with a frequency in the dataset.

Let M be the set of maximal frequent itemsets without specifying the frequency of itemsets.

Let M_c be the set of maximal frequent itemsets each of which is associated with a frequency in the dataset.

The following shows six transactions with four items. Each row corresponds to a transaction where 1 corresponds to a presence of an item and 0 corresponds to an absence.

	A	B	C	D
1	0	0	1	1
2	1	1	0	0
3	0	0	1	1
4	1	0	1	1
5	1	0	0	0
6	0	0	0	1

Tran	Items
1	C D
2	A B
3	C D
4	A C D
5	A
6	D

Suppose that the support threshold is 2.

(a) (i) What is F_c? (ii) What is C_c? (iii) What is M_c? **(5 Marks)**

(b) (i) What are the advantages and the disadvantages of using closed frequent itemsets compared with traditional frequent itemsets? **(5 Marks)**

(ii) What are the advantages and the disadvantages of using closed frequent itemsets compared with maximal frequent itemsets? **(5 Marks)**

(c) Please adapt algorithm FP-growth with the use of the FP-tree to find all closed frequent itemset. Please write down how to adapt algorithm FP-growth and illustrate the adapted algorithm with the above example. **(20 Marks)**

(a)

C_1

Itemset	sup
A	3
B	1
C	3
D	4

L_1

Itemset	sup
A	3
C	3
D	4

C_2

Itemset
AC
AD
CD

C_2

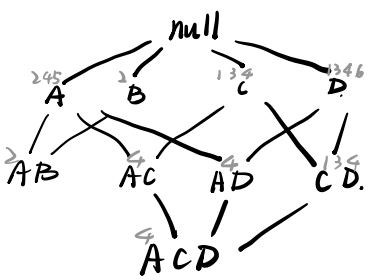
Itemset	Sup
AC	1
AD	1
CD	3

L_2

Itemset	sup
CD	3

(i) $F_c = \{<\{A\}, 3>, <\{C\}, 3>, <\{D\}, 4>, <\{C, D\}, 3>$

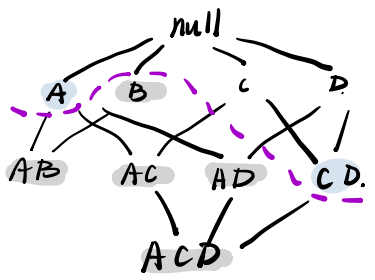
(ii)



$C = \{\{A\}, \{D\}, \{C, D\}\}$

$C_c = \{<\{A\}, 3>, <\{D\}, 4>, <\{C, D\}, 3>\}$

(iii)



$M = \{\{A\}, \{C,D\}\}$

$M_c = \{<\{A\}, 3>, <\{C,D\}, 3>\}$

(b) (i) Compared with traditional frequent itemsets

advantages: Closed frequent itemsets can reduce redundancy and computing is more efficient compared to traditional frequent itemsets, leading to faster results.

disadvantages: Using closed frequent itemsets may not capture all the information from the frequent itemsets and lose some useful frequent itemsets that are proper supersets and so on.

ii) Compared with maximal frequent itemsets

advantages: Maximal frequent itemsets don't contain information about the support of their subsets and therefore an additional scan of transactions is required. On the other hand, closed frequent itemsets don't require this additional scan as they already capture the necessary support information.

disadvantages: Maximal frequent itemsets can inform the small set of itemsets that can derive all frequent itemsets but closed frequent itemsets can't. And implementation of closed frequent itemsets is relatively more complex.

And since maximal frequent itemset is also a part of closed itemset, the advantage and disadvantage of using closed frequent itemset instead of maximal frequent itemset is similar to the answer in (b)(i).

(c) ① step 1: deduce the ordered frequent items

step 2: construct the FP-tree

step 3: construct the FP-conditional tree.

step 4: mine frequent patterns, identify closed itemsets for pruning and determine the frequent patterns.

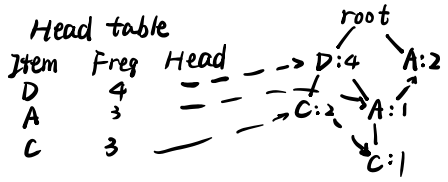
② L_1

Itemset	Frequent
D	4
A	3
C	3

Tran	Items	order frequent Items
1	C D	DC
2	A B	A
3	C D	DC
4	A C D	DAC
5	A	A
6	D	D

③ Head table

Item	Freq	Head
D	4	
A	3	
C	3	



Cond FP-tree on C:3.

{(D:2, C:2), (D:1, A:1, C:1)} → {(D:2, C:2), (D:1, C:1)} → root → D:3

Cond FP-tree on A:3

{(A:2), (D:1, A:1)} → {(A:3)} → root

Cond FP-tree on D:4

{(D:4)} → root

Item	cond pattern-base	Cond FP-tree.	Freq pattern.	closed Freq pattern
C	{{D:2}, {D, A:1}}	<D:3>	{D, C:3}	{D, C:3}
A	{{D:1}}		{A:3}	{A:3}
D			{D:3}	{D:3}

∴ {{C, D}, {A}, {D}}

Q4 [35 Marks]

A GSP Example: Suppose now we have 5 events: 'Upload Songs', 'Add Tags', 'Share', 'Listen' and 'Commet'. Let min-support be 40%. The sequence database of a Music Platform is shown in following table:

Object	Sequence
A	<{ 'Upload Songs', 'Add Tags'}>
B	<{ 'Upload Songs', 'Share'}>
C	<{ 'Upload Songs'}, { 'Share', 'Listen'}>
D	<{ 'Upload Songs'}, { 'Upload Songs', 'Add Tags'}, {'Listen'}>
E	< {'Listen'}, { 'Add Tags', 'Comment'}, { 'Share', 'Listen'}>

Please answer the following questions:

(a) Make the first pass over the sequence database to yield all the 1-element **frequent** sequences and what is the corresponding support? **(5 Marks)**

(b) Based on (a), do the 2-sequences Candidate Generation and Candidate Pruning. **(10 Marks)**

(c) What is the **frequent** 2-sequences based on the results of (b)? **(5 Marks)**

(d) Based on (c), do the 3-sequences Candidate Generation and Candidate Pruning. When a sequence should be pruned, you need to explain why. **(10 Marks)**

(e) What is the **frequent** 3-sequences based on the results of (d)? Please calculate the support. **(5 Marks)**

Remember: For frequent k-sequences, the support >= min-support

(a) Candidate 1-sequences are:

<{'Upload Songs'}>, <{'Add Tags'}>, <{'Share'}>, <{'listen'}>, <{'Comment'}>

80% 60% 60% 60% 20%

∴ all the 1-element frequent sequences are: <{'Upload Songs'}>, <{'Add Tags'}>, <{'Share'}>, <{'listen'}>.

(b) Candidate 2-sequences are:

<{'Upload Songs', 'Add Tags'}>, <{'Upload Songs', 'Share'}>, <{'Upload Songs', 'listen'}>,

<{'Add Tags', 'Share'}>, <{'Add Tags', 'listen'}>, <{'Share', 'listen'}>,

<{'Upload Songs'}, {'Upload Songs'}>, <{'Upload Songs'}, {'Add Tags'}>,

<{'Upload Songs'}, {'Share'}>, <{'Upload Songs'}, {'listen'}>,

<{'Add Tags'}, {'Upload Songs'}>, <{'Add Tags'}, {'Add Tags'}>,

<{'Add Tags'}, {'Share'}>, <{'Add Tags'}, {'listen'}>,

<{'Share'}, {'Upload Songs'}>, <{'Share'}, {'Add Tags'}>,

<{'Share'}, {'Share'}>, <{'Share'}, {'listen'}>,

<{'listen'}, {'Upload Songs'}>, <{'listen'}, {'Add Tags'}>,

<{'listen'}, {'Share'}>, <{'listen'}, {'listen'}>, 22 sequences

After candidate pruning,

<{'Upload Songs', 'Add Tags'}>, <{'Upload Songs', 'Share'}>, <{'Upload Songs', 'listen'}>,

<{'Add Tags', 'Share'}>, <{'Add Tags', 'listen'}>, <{'Share', 'listen'}>,

<{'Upload Songs'}, {'Upload Songs'}>, <{'Upload Songs'}, {'Add Tags'}>,

<{'Upload Songs'}, {'Share'}>, <{'Upload Songs'}, {'listen'}>,

<{'Add Tags'}, {'Upload Songs'}>, <{'Add Tags'}, {'Add Tags'}>,

<{'Add Tags'}, {'Share'}>, <{'Add Tags'}, {'listen'}>,

<{'Share'}, {'Upload Songs'}>, <{'Share'}, {'Add Tags'}>,

<{'Share'}, {'Share'}>, <{'Share'}, {'listen'}>,

<{'listen'}, {'Upload Songs'}>, <{'listen'}, {'Add Tags'}>,

<{'listen'}, {'Share'}>, <{'listen'}, {'listen'}>,

(c) After candidate elimination, the remaining frequent 2-sequences are:

<{'Upload Songs', 'Add Tags'}> (support=0.4)

<{'share', 'listen'}> (support=0.4)

<{'Upload Songs'}, {'listen'}>. (support=0.4)

<{'Add Tags'}, {'listen'}> (support=0.4)

(d) Candidate 3-sequences :

<{'Upload Songs', 'Add Tags'}, {'listen'}>.

generated from <{'Upload Songs', 'Add Tags'}> and <{'Add Tags'}, {'listen'}>

And should not be pruned because all 2-subsequences

<{'Upload Songs'}, {'listen'}>. <{'Add Tags', {'listen'}> are frequent

<{'Upload Songs'}, {'share', 'listen'}>

generated from <{'Upload Songs'}, {listen}> and <{'share', 'listen'}>

and should be pruned because <{'Upload Songs'}, {'share'}> is not frequent.

<{'Add Tags'}, {'share', 'listen'}>

generated from <{'Add Tags'}, {listen}> and <{'share', 'listen'}>

and should be pruned because <{'Add Tags'}, {'share'}> is not frequent.

So. aftering pruning, the remaining 3-sequence:

<{'Upload Songs', 'Add Tags'}, {'listen'}>

(e) <{'Upload Songs', 'Add Tags'}, {'listen'}>. support = 0.2 < 0.4,

should be eliminated.

Thus, there are no 3-sequences left.