DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

(Fall Semester 2023)

Homework 1

Deadline: 4 Oct 2023 11:59pm

(Please hand in via Canvas.)

Full Mark: 100 Marks

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

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

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 $\mathrm{F_c}$ 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, $\{\mathrm{I}_1\}$ with frequency 4, $\{\mathrm{I}_2\}$ with frequency 5, and $\{\mathrm{I}_1, \mathrm{I}_2\}$ with frequency 3, F = $\{\{\mathrm{I}_1\}, \{\mathrm{I}_2\}, \{\mathrm{I}_1,\mathrm{I}_{12}\}\}$ and Fc = {<$\{\mathrm{I}_1\}$, 4>, <{II_2}, 5>, <$\{\mathrm{I}_1, \mathrm{I}_2\}$, 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
0	0	1	1
1	1	0	0
0	0	1	1
1	0	1	1
1	0	0	0
0	0	0	1

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)**

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