TP2 : Frequent Itemset Mining

Exercise 1

Let $\mathcal{D}_1$ be a transactional database represented in the horizontal format $\mathcal{H}_{\mathcal{D}_1}$ as follows :

Trans.			**Items**			
t_1		B	C	D		
t_2	A	B	C		E	
t_3	A	B	C	D		F
t_4				D	E	
t_5	A	B				
t_6	A		C		E	F
t_7	A	B			E	F
t_8				D		F
t_9			C		E	
t_{10}	A	B				F

Question 1 • Provide the vertical representation $\mathcal{V}_{\mathcal{D}_1}$ and the matrix representation $\mathcal{M}_{\mathcal{D}_1}$ of $\mathcal{D}_1$.

Question 2 • Calculate the support, absolute frequency, and relative frequency of the following itemsets :

$$L = \{ACD, CE, BCE, ABCE, E, D, BC, F, CDF, EF\}.$$

Question 3 • Identify the frequent itemsets with minimum support values $\alpha \in \{5, 6, 7, 8, 9, 10\}$.

Question 4 • Provide an example of two comparable itemsets and two non-comparable itemsets.

Exercise 2

Question 1 • Write a proof for the anti-monotone property of frequent itemsets.

Question 2 • Write a proof for the Apriori property.

Exercise 3

Let $\mathcal{D}_2$ be a transactional database as follows :

Trans.	Items				
t_1	A		C	D	
t_2		B	C		E
t_3	A	B	C		E
t_4		B			E
t_5	A	B	C		E
t_6		B	C		E

Question 1 • Run the Apriori algorithm on $\mathcal{D}_2$ with a minimum support $\alpha = 3$, without using the canonical operator κ.

Question 2 • Run the Apriori algorithm on $\mathcal{D}_2$ with a minimum support $\alpha = 3$, using the `child` operator based on a lexicographical order `lex`.

Question 3 • Implement the Apriori algorithm in Java with and without the `child+lex` operator. Compare the performance of the two versions on the datasets provided in .\`DataSets`\.

Question 4 • Propose an algorithm with a bottom-up exploration approach to extract the set of frequent itemsets. Implement it and compare its performance with the Apriori algorithm.

Question 5 • Revise the Apriori algorithm to extract only frequent itemsets with a size greater than a specified value *size*. Implement this modified version.

Exercise 4

Let the set of maximal itemsets M_α be as follows : $M_\alpha = \{ABC^3, DE^2, EF^5\}$

Question 1 • Provide the list of frequent itemsets.

Let the set of closed itemsets C_α be as follows : $C_\alpha = \{ABC^3, ABE^5, DE^2, EF^5\}$ • Provide the list of frequent itemsets.

Question 2 • Consider now the transactional database $\mathcal{D}_2$ given before. Determine the sets of maximal and closed frequent itemsets with a minimum support $\alpha = 3$.

Exercise 5

Question 3 • Run the LCM algorithm on $\mathcal{D}_1$ with a minimum support threshold $\alpha = 3$.

Question 4 • Implement the LCM algorithm in Java. Test the performance of your implementation on the datasets provided in .\`DataSets`\.

Exercise 6

Consider the following query :

$$Q : frequent(P) \wedge closed(P) \wedge maxSize_{ub}(P)$$

with two interpretations :

1. Mine all frequent closed itemsets that additionally have a size less than or equal to *ub*.
2. Mine all frequent itemsets of size less than or equal to *ub* that additionally have the property of being closed.

Question 1 • Provide the set of solutions for Q under both interpretations on the dataset $\mathcal{D}_1$ with a minimum support threshold $\theta = 3$.

Question 2 • What is the correct semantic of this query ? Explain your reasoning.