Exercises Frequent Pattern Mining 2018

Exercise 1: Frequent Item Set Mining

Given are the following eight transactions on items $\{A, B, C, D, E, F\}$:

tid	items
1	ABC
2	BCD
3	CDE
4	BC
5	CD
6	$ABCD$
7	ABD
8	EF

(a) Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. Clearly indicate the steps of the algorithm, and the pruning that is performed.

(b) Use the Apriori-close algorithm to compute all *closed* frequent item sets, and their support, with minimum support 2. Clearly indicate the steps of the algorithm, and the extra pruning that is performed.

(c) Give the maximal frequent item sets.

(d) Compute the confidence and the lift of the rule $A \rightarrow C$. Do you find this rule interesting?

Exercise 2: Frequent Sequence Mining

Consider the following database of travel sequences for one working week of some anonymous person:

day	sequence
Mon	AUHUA
Tue	AUHUB
Wed	BA
Thu	AUHUA
Fri	AUB

The meaning of the symbols is:

- A: **A**msterdam Central Station
- U: **U**trecht Central Station
- H: Utrecht Uit**H**of Busstop
- B: **B**reda Trainstation

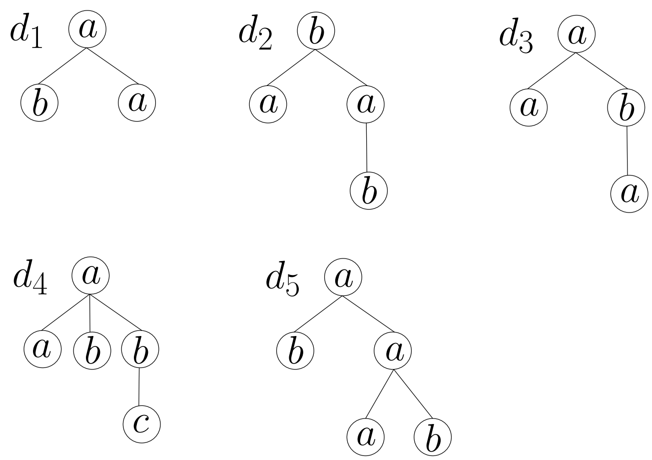
(a) Use the GSP algorithm to find all frequent sequences with minsup=3. For each level, make a table listing all candidates and their support. Also indicate whether a candidate sequence is frequent. Pre-candidates that have an infrequent subsequence should not be listed in the table!

(b) A frequent sequence is maximal frequent if it doesn't have a frequent super-sequence. Which of the frequent sequences are maximal?

(c) A frequent sequence is closed frequent if it doesn't have a super-sequence with the same support. Which of the frequent sequences are closed?

Exercise 3: Frequent Tree Mining

Consider the following database of ordered labeled trees:



We use the following string representation of an ordered labeled tree: list the labels according to the pre-order traversal of the tree, and use the special symbol $\uparrow$ to indicate we go up one level in the tree. For example, the string representation of d_4 is: $aa \uparrow b \uparrow bc$.

Answer the following questions:

(a) Is $aa \uparrow c$ an induced subtree of d_4?
If yes, give the corresponding matching function(s).

(b) Is $aa \uparrow c$ an embedded subtree of d_4?
If yes, give the corresponding matching function(s).

(c) Is d_1 an induced subtree of d_4? If yes, give the corresponding matching function(s).

(d) Is d_1 an embedded subtree of d_4? If yes, give the corresponding matching function(s).

(e) Is d_1 an induced subtree of d_5? If yes, give the corresponding matching function(s).

(f) Is d_1 an embedded subtree of d_5? If yes, give the corresponding matching function(s).

(g) Consider the ordered labeled tree $ab \uparrow bb \uparrow\uparrow bb$. How many times does $ab \uparrow b$ occur as an embedded subtree? Give the corresponding matching functions.

(h) Consider the ordered labeled tree $ab \uparrow bb \uparrow\uparrow bb$. How many times does $ab \uparrow b$ occur as an induced subtree? Give the corresponding matching functions. Also give the FREQT right-most occurrence list (RMO list) for $ab \uparrow b$ in $ab \uparrow bb \uparrow\uparrow bb$.

Exercise 4: Anti-monotonicity

Consider an alternative sequence mining scenario, where we have just a single data sequence. In this scenario, the support of a pattern sequence is equal to the number of distinct occurrences of the pattern sequence in the data sequence. Two occurrences are considered distinct if they correspond to mapping functions ϕ_1 and ϕ_2, where $\phi_1(i) \neq \phi_2(i)$ for some position i in the pattern sequence.

Do we have the anti-monotonicity property between support and the subsequence relationship in this scenario? Explain.

Can you think of another reasonable definition of "distinct occurrence"? Do we have the anti-monotonicity property in that case?

Exercise 5: Transitivity of the subsequence relation

To show that the subsequence relation is anti-monotone with respect to support, it suffices to show that the subsequence relation is transitive. Explain why this is so.

Let $\mathbf{q}, \mathbf{r}$, and $\mathbf{s}$ be arbitrary sequences over some set of labels Σ.
Show that the subsequence relation is transitive: if $\mathbf{q} \subseteq \mathbf{r}$, and $\mathbf{r} \subseteq \mathbf{s}$, then $\mathbf{q} \subseteq \mathbf{s}$.

For your convenience, we recall the definition of the subsequence relation: we say $\mathbf{r} = r_1 r_2 \ldots r_m$ is a subsequence of $\mathbf{s} = s_1 s_2 \ldots s_n$, denoted $\mathbf{r} \subseteq \mathbf{s}$, if there exists a one-to-one mapping $\phi : [1, m] \rightarrow [1, n]$, such that

1. $\mathbf{r}[i] = \mathbf{s}[\phi(i)]$, and
2. $i < j \Rightarrow \phi(i) < \phi(j)$.

Exercise 6: Variations on a theme

Consider data of supermarket customers with a loyalty card. Now we can track a customer's purchases through time. Define an item set sequence as a sequence $S = (X_1 X_2 \ldots X_k)$ where each X_i is an item set. We have a database $D = \{S^1, S^2, \ldots, S^N\}$ of such sequences, one for each customer with a loyalty card.

Give plausible definitions for the subsequence relation and for the support of a sequence. Verify that the subsequence relation has the anti-monotonicity property with respect to support.