Data Mining 2: Frequent Pattern Mining Summer 2018
Christian Borgelt from 25.06.2018

Exercise Sheet 8

Exercise 28 Closed and Maximal Item Sets

a) Suppose we tried to find all closed (or all maximal) (frequent) item sets by first finding all frequent item sets and then filtering the result according to the defining conditions for closed (or maximal) item sets. Is this a good idea? Justify your answer!

b) What do the notions *head* and *tail* of a search tree node (or a subproblem) mean? Is any of them related to the *prefix* of a subproblem?

c) Given that we use a global item order, is there a (fixed) relationship between the items in the head and those in the tail?

d) Does the relation $head \cup tail = B$ hold for all search tree nodes? If no, are there any search tree nodes for which this holds or any specific conditions that must be satisfied so that this holds?

Exercise 29 Excluded Items

a) What is the set of *excluded items* (relative to a search tree node or subproblem)? How can one characterize this set based on the subproblem splits we carry out in the divide-and-conquer algorithm? What could be called, in analogy, the set of *included items*?

b) Suppose the divide-and-conquer scheme for finding frequent item sets
is extended with perfect extension pruning,
so that subproblems are described by triplets $S = (T, P, X)$.
Is $P \cup X$ always a closed item set? Is it always a maximal item set?
Justify your answer!

c) Is the following statement true? If the tail of a search tree node is empty, its head is a maximal item set. Justify your answer!

d) Is the following statement true? If no item in the tail of a search tree node has the same support as the head, the head is a closed item set. Justify your answer!

Exercise 30 Closed and Maximal Item Sets

a) How can one check the defining condition of closed item sets with a *horizontal* transaction (database) representation?

b) How can one check the defining condition of closed item sets with a *vertical* transaction (database) representation?

c) What is an alternative to checking the defining condition?
How is the check performed in this alternative approach?
What are possibilities (e.g. data structures) to implement this alternative?

Exercise 31 Closed and Maximal Item Sets: Pruning

a) In the preceding exercises we considered how one can ascertain whether an item set is closed (or maximal) and thus how we can avoid reporting non-closed (or non-maximal) item sets. Is this simple filtering the only advantage? Or can we simplify/prune the search with the help of these methods as well? If yes, how?

b) How is perfect extension pruning applied in the search for closed item sets? Why is it guaranteed that we do not lose closed item sets due to such pruning?

c) What is meant by "head union tail pruning" for finding maximal item sets (two variants)?

Additional Exercise Alternatives to Support

a) Find some measure(s) other than support on the partial order $(2^B, \subseteq)$ that is/are anti-monotone (or downward closed)!

b) We define the *carrier* $L_T(I)$ of an item set I w.r.t. a transaction database T as

$$\begin{aligned} L_T(I) &= \{k \in \{1, \ldots, n\} \mid I \cap t_k \neq \emptyset\} \\ &= \{k \in \{1, \ldots, n\} \mid \exists i \in I \colon i \in t_k\} \\ &= \textstyle\bigcup_{i \in I} K_T(\{i\}). \end{aligned}$$

The *extent* $r_T(I)$ of an item set I w.r.t. a transaction database T is the size of its carrier, that is, $r_T(I) = |L_T(I)|$. Is the extent anti-monotone (or downward closed)? If not, what behavior does it exhibit?

c) Could one combine support and extent to obtain a more meaningful measure for the association of the items in an item set? (Hint: Look up the measure called *Jaccard index* and consider how it could be generalized.)