581550 Data mining — tietämyksen muodostaminen
Autumn 2002
Hannu Toivonen

Exercises 4 (due Oct 7–Oct 11)

1. Study *exact database rules* as "frequent" patterns (Section 7.3.3 in the course material).

   Explain what the specialization relation $\preceq$ is and how it works. Explain the relation $\sqsubseteq$. Use examples.

2. Explain Theorem 7.12 and its proof.

3. Consider the problem of finding common strings in given lines of text. For instance, string "`a da`" occurs on both lines

   ```
   abracadabra da cabra
   ta daa
   ```

   but string `abra` only on the first one. Formulate the knowledge discovery task. In particular, specify

   - $\mathcal{P}$: class of patterns
   - $q$: selection criterion
   - $\preceq$: specialization relation

4. Give a simple simulated example of Algorithm 7.6 being applied to some data with your specifications in task 3 above.

5. Consider the class of frequent sets over $R = \{A, \ldots, E\}$ and a set of patterns $\mathcal{S} = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B\}, \{A, D\}, \{A, E\}, \{B, D\}, \{D, E\}, \{A, B, D\}, \{A, D, E\}\}$.

   - What is the positive border $\mathcal{B}d^+(\mathcal{S})$?
   - What is the negative border $\mathcal{B}d^-(\mathcal{S})$?

6. Consider the pattern class of frequent sets, and the connection between the border of a theory $\mathcal{T}$ and hypergraph transversals. Assume $R = \{A, \ldots, E\}$ and $\mathcal{B}d^+(\mathcal{T}) = \{\mathcal{A}, \mathcal{C}\}, \{\mathcal{A}, \mathcal{D}, \mathcal{E}\}, \{\mathcal{B}, \mathcal{C}\}, \{\mathcal{B}, \mathcal{D}, \mathcal{E}\}$.

   What is the corresponding hypergraph?

   Obtain $\mathcal{B}d^-(\mathcal{T})$ by computing the minimal transversals of this graph.