581550 Data mining — tietämyksen muodostaminen
Autumn 2002
Hannu Toivonen

Exercises 5 (due Oct 15–Oct 18)

1. Find the frequent generators ($min\_fr = 2/5$) in the following database over $R = \{A, \ldots, E\}$. Use the modified Apriori algorithm (e.g. the one given on the "closed sets" slides).

A, D, E
A, D
A, C, D, E
C, E
A, C, D

   – Walk through the algorithm. Give candidates, frequent generators, and the negative border for each level.

   – What are the corresponding closed sets?

2. Let

   – $R = \{A, \ldots, E\}$

   – $\mathcal{C}\ell = \{(B, 0.07), (C, 0.33), (E, 0.31), (AD, 0.45), (CE, 0.23), (ACD, 0.22), (ADE, 0.21), (ACDE, 0.13), (ABCDE, 0)\}$

   – $\mathcal{G}en = \{(A, 0.45), (B, 0.07), (C, 0.33), (D, 0.45), (E, 0.31), (AC, 0.22), (AE, 0.21), (CD, 0.22), (CE, 0.23), (DE, 0.21), (ACE, 0.13), (CDE, 0.13)\}$

   where a pair such as $(ACE, 0.13)$ denotes that the frequency of set $\{A, C, E\}$ is 0.13. Determine the frequencies of sets $\{A, C\}$, $\{B, C\}$, and $\{A, C, D\}$

   – using generators only

   – using closed sets only

3. Prove that $fr(X) = \max\{fr(Y) \mid Y \in \mathcal{C}\ell \text{ and } X \subseteq Y\}$

4. Construct the FP-tree (including node links and counts) for the dataset of task 1.

5.–6. Use CLOSET algorithm to find closed sets in the FP-tree of task 4.