T-61.5060 Algorithmic methods in data mining

Exercises 4, October 11, 2007

1. The file eventsequence.txt available from the web site of the course contains an event sequence with 1000 events; the event types are coded as integers 1, 2, 3 and 4. The first column is the occurrence time and the second is the event type. Do the event types seem randomly distributed?

2. Given two event types A and B, let $d(A, B)$ be the average distance from an occurrence of A to the next occurrence of B, and $s(A, B)$ the standard deviation of the distance from an occurrence of A to the next occurrence of B. Given a long event sequence with many different types of events, how would you compute $d(A, B)$ and $s(A, B)$ for all event types A and B?

3. In many cases finding frequent patterns is not sufficient; rather, we want to enforce some aspect of minimality

   Consider a set $\{(x_i, y_i)|i = 1, \ldots, n\}$ of points in the plane. A rectangle $R(a, b, c, d)$ is frequent, if there are at least K points (x_i, y_i) with $a \leq x_i \leq b$ and $c \leq y_i \leq d$. Why is it not a good idea to search for all frequent rectangles?

   A frequent rectangle $R(a, b, c, d)$ is minimal, if there is no frequent rectangle $R(a', b', c', d')$ which is properly contained in $R(a, b, c, d)$ [i.e., $a \leq a' \leq b' \leq b$ and $c \leq c' \leq d' \leq d$, with at least one of the inequalities $a < a'$, $b' < b$, $c < c'$ and $d' < d$ holding].

   Describe a method for finding all minimal frequent rectangles. What is the complexity of your method?