Assignment 4 (Knowledge Discovery and Data Mining)

Number of problems/points: Three problems for total of 100 points
Out: March 19, 2010
Due: March 31, in class.

Problem 1. (40 points)
Consider the following market basket transactions:

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Transaction ID</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0001</td>
<td>{a, d, e}</td>
</tr>
<tr>
<td>1</td>
<td>0024</td>
<td>{a, b, c, e}</td>
</tr>
<tr>
<td>2</td>
<td>0012</td>
<td>{a, b, d, e}</td>
</tr>
<tr>
<td>2</td>
<td>0031</td>
<td>{a, c, d, e}</td>
</tr>
<tr>
<td>3</td>
<td>0015</td>
<td>{b, c, e}</td>
</tr>
<tr>
<td>3</td>
<td>0022</td>
<td>{b, d, e}</td>
</tr>
<tr>
<td>4</td>
<td>0029</td>
<td>{c, d}</td>
</tr>
<tr>
<td>4</td>
<td>0040</td>
<td>{a, b, e}</td>
</tr>
<tr>
<td>5</td>
<td>0033</td>
<td>{a, d, e}</td>
</tr>
<tr>
<td>5</td>
<td>0038</td>
<td>{a, b, e}</td>
</tr>
</tbody>
</table>

(a) Compute the support for itemsets \{e\}, \{b, d\}, and \{b, d, e\} by treating each transaction ID as a market basket.

(b) Use the results in part (a) to compute the confidence for the association rules \{b, d\} \rightarrow \{e\} and \{e\} \rightarrow \{b, d\}. Is confidence a symmetric measure?

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

(d) Use the results in part (c) to compute the confidence for the association rules \{b, d\} \rightarrow \{e\} and \{e\} \rightarrow \{b, d\}.

(e) Suppose \(s_1\) and \(c_1\) are the support and confidence values of an association rule \(r\) when treating each transaction ID as a market basket. Also, let \(s_2\) and \(c_2\) be the support and confidence values of \(r\) when treating each customer ID as a market basket. Discuss whether there are any relationships between \(s_1\) and \(s_2\) or \(c_1\) and \(c_2\).

Problem 2. (10 points)
Propose three topics on which you would like to present in class. These topics should be different from topics already discussed in class. Each topic should be appropriate for a 60 minutes presentation (given on first 3 Wednesdays in April). For each topic list literature that you would use to prepare the lecture. Possible topics to consider (you can also consider different topics):

- Mining spatial data
- Mining time series and sequence data
- Mining spatio-temporal data
- Data stream mining
- Mining graphs
- Anomaly detection
- Text and web mining
- Mining semistructured data
- Mining with constraints
- Frequent pattern mining
- Privacy preserving data mining

**Problem 3. (50 points)**

Write a research proposal for the class project that you plan to perform. Use the following format:

(0) Your name and **e-mail address**
(1) Title;
(2) Objective and Significance;
(3) Background;
(4) Proposed Approach;
(5) References.

The description may not exceed 2 pages in 12 pt style.