Data mining, Spring 2010. Exercises 1, March 30, 2010

1. An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 multiple choice questions with four possible answers each. How would you convert this data into a form suitable for association analysis?

2. Consider the market basket data in the Table 1 below.

Transaction ID	Items bought
1	Milk,Beer,Diapers
2	Bread, Butter, Milk
3	Milk, Diapers, Cookies
4	Bread, Butter, Cookies
5	Beer, Cookies, Diapers
6	Milk, Diapers, Bread, Butter
7	Bread, Butter, Diapers
8	Beer, Diapers
9	Milk, Diapers, Bread, Butter
10	Beer, Cookies

Table 1: Marker basket transactions

   (a) What is the maximum number of association rules that can be extracted from this data (including rules with zero support)?

   (b) What is the maximum size (k) of frequent k-itemsets in this data, assuming $minsup > 0$.

3. From the data of Table 1,

   (a) Find a itemset with 2 or more items that has the largest support.

   (b) Find a pair of items a and b, such that the rules $a \rightarrow b$ and $b \rightarrow a$ have the same confidence.

4. Consider the following set of frequent 3-itemsets: $\{1,2,3\}, \{1,2,4\}$, $\{1,2,5\}, \{1,3,4\}$, $\{1,3,5\}, \{2,3,4\}$, $\{2,3,5\}, \{3,4,5\}$.

Assume that there are only five items in the dataset.

List all candidate 4-itemsets obtained by the $F_{k-1} \times F_1$ candidate generation method.

5. Consider the same set of frequent 3-itemsets as above. List all candidate 4-itemsets obtained by the $F_{k-1} \times F_{k-1}$ candidate generation method.

6. –7. Consider the following set of candidate 3-itemsets: $\{1, 2, 3\}$,$\{1, 2, 6\}$,$\{1, 3, 4\}$, $\{2, 3, 4\}$, $\{2, 4, 5\}$, $\{3, 4, 6\}$, $\{4, 5, 6\}$,

   Construct a hash tree for the itemsets, using a hash function that sends odd-numbered items to the left and even-numbered items to the right.

   A candidate itemset is inserted by hashing each successive item in the candidate and following the appropriate branch in the hash tree.

   Once a leaf node is reached the candidate is inserted according to the following conditions:

   Condition 1: if the depth of leaf equals k (root is on level 0) the candidate is inserted regardless of how many itemsets are already stored at the node.

   Condition 2: if the depth of the node is less than k, then the candidate is inserted as long as there are less than $maxsize = 2$ itemsets already stored at the node.

   Condition 3: if the depth of node is less than k and the number of itemsets in the node is $maxsize = 2$, convert the leaf into an internal node. New leafs are created as the children of the node and the new itemset as well as itemsets previously stored in the node are hashed into the children using the hash function.