CMPT 454 (Spring 2010)
Assignment 4: Advanced Queries and Data Mining

Due: Friday, 2010-04-09, 11:29AM (at the beginning of class)

Instructions on Assignment Submission: Hard copy only!

- Option 1: drop off your assignments in the assignment box (with label **CMPT 454**) in CSIL.
- Option 2: bring your assignments to the class on the due day, and the instructor will collect the assignments **at the beginning of class**.

Please write legibly or typeset your answers using your favorite word processor. Late assignments will not be accepted unless there is a documented medical reason.

Problem 1.

(20 points) [*kd*-Trees]

Figure 1 is a *kd*-tree for twelve 2-dimensional points (the first dimension refers to *age*, and the second dimension refers to *salary* (in thousand)). In what situations, new points would be directed to the following cases?

(1) The block with point (30, 260)?

(2) The block with points (50, 100) and (50, 120)?

(3) The block with points (45, 60) and (50, 75)?

(4) The block with point (25, 60)?

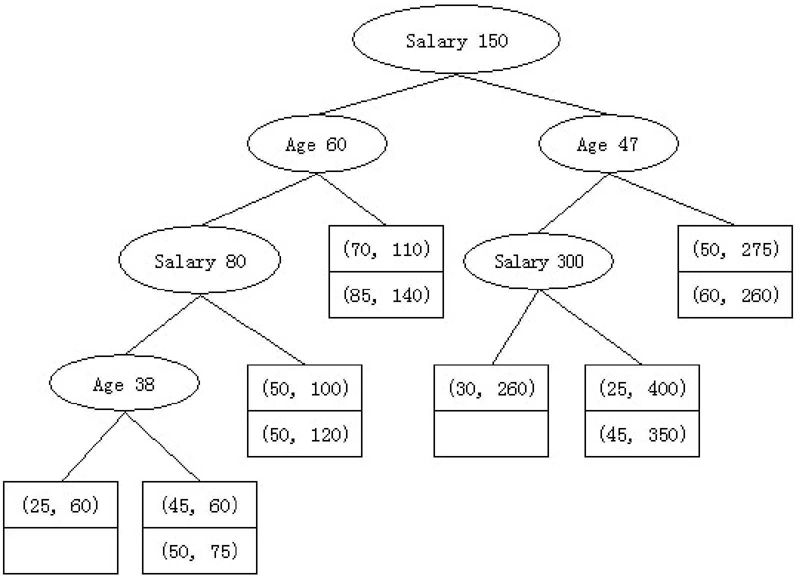

Figure 1: Problem 1.

Problem 2.

(25 points) [Skyline Query]

(1) Suppose we have a set of 2-dimensional data points (representing hotels, the first dimension refers to *price*, and the second dimension refers to *distance to beach*) listed as following: $p_1(5,17)$, $p_2(4,16)$, $p_3(12,12)$,$p_4(11,20)$, $p_5(8,8)$, $p_6(16,12)$, $p_7(17,9)$, $p_8(10,18)$, $p_9(20,4)$, $p_{10}(21,5)$. The smaller the price and the distance to beach, the better the hotel is. Please find all the skyline points.

(2) Suppose we have an extended SQL operator *SKYLINE OF* to extend SQL's SELECT statement. For a relation *Hotels(price, distance)* which contains hotels in Vancouver, we have the following extended SQL query to find all the skyline hotels.

SELECT $*$
FROM Hotels
SKYLINE OF price MIN, distance MIN

Actually the SKYLINE OF operator can be implemented using the traditional SQL's SELECT statement you have learned from CMPT 354. Please provide an equivalent SQL query (without the SKYLINE OF operator) to find all the skyline hotels.

Problem 3.

(25 points) [Frequent Itemsets and Association Rules]

Suppose we are given the following 8 "market baskets":

B_1 = {milk, coke, beer}	B_5 = {milk, pepsi, beer}
B_2 = {milk, pepsi, juice}	B_6 = {milk, beer, juice, pepsi}
B_3 = {milk, beer}	B_7 = {coke, beer, juice}
B_4 = {coke, juice}	B_8 = {beer, pepsi}

Please answer the following several questions.

(1) What is the *support* of the itemset {beer, juice}?

(2) What is the *support* of the itemset {coke, pepsi}?

(3) What is the *confidence* of milk given beer (i.e., of the association rule {beer} $\Rightarrow$ milk)?

(4) What is the *confidence* of coke, given beer and juice?

(5) If the *support* threshold is 3, which pairs of items (two items) are frequent?

Problem 4.

(30 points) [R-Trees and NN Search]

This is an open algorithm-design problem. Given two sets of points S and T, for a point $s \in S$, a point $t \in T$ is the *nearest neighbor of s in T* if $dist(s,t) \leq dist(s,t')$ for any $t' \in T$. Suppose S and T are indexed individually using two R-trees RT_S and RT_T, respectively. Please describe an algorithm as efficient as you can to compute the nearest neighbors in T for all points in S.