Written Assignment #2
CIS5930: Advanced Topics in Data Management
Fall 2009

Assigned: Tuesday, Oct 21, 2008; Due: In Class on Tuesday, Nov 4, 2008.

Problem 1. [40 points]

A. Describe an algorithm for finding the k-furthest neighbors using R-tree. Please use graph to illustrate your idea when necessary.

B. Given two data sets P and Q, for any point $q \in Q$, we would like to find all the points in P that take q as their nearest neighbors comparing to all other points in Q, i.e., Let $NN(p, Q)$ denote the nearest neighbor of a point p in a dataset Q, we are interested at finding out the set $\{p|p \in P \bigwedge NN(p, Q) = q\}$. Describe an algorithm based on R-tree to achieve this. Please use graph to illustrate your idea when necessary. Any types of brute-force search based algorithms will not be credited.

Problem 2. [30pts]

Given two multiset A and B, and their corresponding FM sketch $S(A)$ and $S(B)$. Please answer the followings:

1. (10 pts) Please design a method to estimate the distinct number of elements in $A \bigcap B$ without building the FM sketch for it from the scratch. Explain your answer.
2. (10 pts) We use bit-wise AND operation on $S(A)$ and $S(B)$, the resulting sketch represents the sketch for $A \bigcap B$. Is this correct? If Yes, please prove it. If No, please explain why it is wrong.
3. (10 pts) Please design an algorithm to compress the FM-sketch. Using some examples to illustrate your idea. Keep in mind, by compressing the FM-sketch, you must be able to decompress it without any error.

Problem 3. [30pts]

The current version utilization (CVU) for MVB-Tree is defined as the following: for any given page, CVU is the percentage of live entries. Similarly, the historical version utilization (HVU) for MVB-Tree is defined as: for any given page, HVU is the percentage of the dead/historical entries. Then, finally, the overall utilization (OU) for MVB-Tree is defined similarly as that in B+-Tree: for any page, the total number of entries divided by the total number of potential entries that one page can hold.

Please analyze the lower bound and upper bound for CVU, HVU and OU for MVB-Tree. Explain your answer.