Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

MSBD6000J Spatial and Multimedia Databases

2021-2022 Fall

Version 1.0

Assignment 2 [*Total: 30 marks*]

Due date: 5 pm 23 Nov 2021
HKUST Canvas online submission only.

We will continue to use the POI dataset *D* used in Assignment 1.

Task 1 [*12 marks*] R*-tree implementation.

(1) [*5 marks*] Write a program to create an R*-tree index for point data in-memory. The fan-out of the tree should be *d* (i.e., a non-leaf node can have a maximum of *d* MBRs/subtrees), and each leaf node can contain a maximum of *n* points (i.e., the bucket size is *n*), where both *d* and *n* are user-given parameters. You can implement your program by looking at or using any code online (please make sure that the code is correct and suitable for this assignment, and you do understand the code! The source of the code must be acknowledged in your report).

(2) [*3 marks*] Provide a concise outline of the algorithm you implement, with sufficient plain English comments such that your code can be easily understood by other people.

(3) [*4 marks*] Using the program developed in Task 1 to create an R*-tree index for the POI dataset *D*, and report the following statistics for *n*=128 and 256 and *d* = 2 and 5 (i.e., 4 cases):

   a. [*1 mark*] the height of your R*-tree index.
   b. [*1 mark*] the numbers of non-leaf and leaf nodes.
   c. [*1 mark*] space utilisation for leaf nodes (i.e., the number and percentage of buckets in the following utilisation ranges: less than 20%, 20~80%, more than 80%).
   d. [*1 mark*] sub-tree overlapping among the non-leaf nodes (i.e., the total number of MBR pairs which belong to the same non-leaf node and overlap).

Task 2 [*14 marks*] NN search.

(1) [*7 marks*] Write a program that can find the nearest neighbour of a given query point based on the R*-tree implemented in Task 1 using the algorithm with the three pruning rules discussed in this course. We use L_2 distance in this task.

   a) Input: a query point *q*.
   b) Output: a point *p* in *D* that is the nearest neighbour of *q*.

(2) [*3 marks*] Please describe concisely the algorithm you implement in your program, with sufficient plain English comments such that your code can be easily understood by other people.

(3) [*4 marks*] Run the nearest neighbour program you implemented above, and report, for *n*=128 and 156, *d* = 2 and 5, and 10 random query points, the following statistics for each experiment (there are a total of 2×2×10=40 sets of experiments):

   a. [*1 marks*] The nearest neighbour result *p* and its L_2 distance to *q*.
   b. [*1 marks*] The number of R*-tree nodes visited.

c. [*1 marks*] The number of points where the actual distance to *q* is calculated.
d. [*1 marks*] The number of MBRs that have been pruned out using each of the three pruning rules.

Task 3 [*4 marks*] You need to prepare a report for this assignment for the above results and discussions. The report should be no more than 6 pages (using this document as the template). You may wish to use tables to show your results. Your report should be well structured and carefully written. Your marks will be given based on readability, correctness, completeness, insightfulness, and conciseness of the discussions.

Note: *In this assignment, you can use any programming language of your choice. No programming support will be provided in this course. No DBMS is needed. You will load the entire dataset into memory and perform all operations required in memory including the R*-trees.*

Warning: *This is an individual assignment. Collusion can be easily detected by software tools. Plagiarism will not be tolerated at HKUST. Please refer to Student Conduct and Academic Integrity regulations. If you are unclear about what level of discussions and help you can get for this assignment, please talk to the lecturer or the tutor.*