Boston University
CAS CS 562: Advanced Database Applications

Homework #2
Fall 2020

Due Date: April 2, 2020 at 11:59PM, please submit a PDF file using Gradescope.

Problem 1

Consider the following temporal evolution of a set of records. The first value is the time of the operation, the second value the object id, the third value is the value that is used for indexing and searching, and the last value is the operation. The operation can be either an insertion (i) or a deletion (d).

(time, oid, value, operation)

(1, 1, 34, i)	(11, 1, 34, d)	(21, 21, 43, i)	(31, 20, 122, d)
(2, 2, 23, i)	(12, 12, 27, i)	(22, 22, 32, i)	(32, 9, 29, d)
(3, 3, 4, i)	(13, 2, 23, d)	(23, 23, 44, i)	(33, 33, 81, i)
(4, 4, 56, i)	(14, 14, 55, i)	(24, 10, 6, d)	(34, 34, 21, i)
(5, 5, 7, i)	(15, 5, 7, d)	(25, 14, 55, d)	(35, 35, 57, i)
(6, 6, 24, i)	(16, 6, 24, d)	(26, 26, 72, i)	(36, 36, 16, i)
(7, 7, 3, i)	(17, 7, 3, d)	(27, 27, 13, i)	(37, 27, 13, d)
(8, 8, 49, i)	(18, 18, 19, i)	(28, 18, 19, d)	(38, 30, 50, d)
(9, 9, 29, i)	(19, 19, 8, i)	(29, 19, 8, d)	(39, 12, 27, d)
(10, 10, 6, i)	(20, 20, 122, i)	(30, 30, 50, i)	(40, 40, 38, i)

(a) Create a Snapshot Index on the evolution given above using pages with capacity b=4 entries (number of entries per page(block)), and utilization factor u = 0.5. Show the index after each 10 operations (after 10, 20,30, and 40 operations).

(b) Create a MVB-Tree on the **value attribute** of the above evolution using pages with capacity b=6 entries (number of entries per page(block)), d =2, and $\varepsilon = 0.5$. Show the index after each 10 operations (i.e., after the 10th, 20th, 30th, and 40th operations).

(c) Show how to answer to the following queries using the indices that you created in (a) and (b). Give the pages that you need to retrieve to answer these queries:

1) Timeslice query: Give all the alive records at time tq1 = 30.

2) Range timeslice query: Give all the alive records at time tq3 = 15 that have values between 25 and 35.

Problem 2

Consider the following Time-Interval query: Given a time period T = [t_s, t_e], find all the records that have lifespan that intersects T. That is, find all the records that were alive at least one time instant during the period T. Provide an algorithm and explain what is the cost of this algorithm in terms of number of I/Os. Hint: Read the Snapshot Index paper.

Problem 3

Consider the following 2-dimensional time series:
A = [(2,3), (3,4), (4,5), (4,4), (5,5)] and B=[(2,2),(3,3),(15,2),(4,3),(4,5),(6,8)].

Compute the DTW and the LCSS distances between the two time series. For LCSS use $\varepsilon = 1.5$. To compute the distance between two 2-d points x=(x1, x2) and y=(y1, y2) you can use the L1 distance (L1(x, y) = |x1-y1| +|x2-y2|).

Problem 4

(a) Consider the GEMINI approach to index multimedia and time series data. What happens when the Lower bounding Lemma of the transformation to a lower dimensional space or feature extraction does not hold? Namely, when the $D_{feature}(F(x), F(y)) <= D(x, y)$ does not hold.

(b) Assume that you can prove that $D_{feature}(F(x), F(y)) <= 2\ D(x, y)$. How you will modify the GEMINI approach for RangeQuery and K_NNQuery algorithms in order to guarantee that the complete (correct) answer will be provided at the end?