Homework #5 Answers
due Thursday, November 18 , 2010 at 2 pm

Database Systems, CSCI-4380-01

Each student must work on this homework alone.

1 Homework Description

See attached for question 1 and 2.

Question 3. Suppose you are given a relation with 500 million tuples. Suppose you want to create an index on attributes (X,Y) where each index node may contain at most 500 entries (key value, pointer pairs).

- How many nodes total are in this B-tree if all the nodes (except for the root) contain about 250 entries?

  This means there are $500 * 10^6/250 = 2x10^6$ nodes at the leaf level.

  $2x10^6/250 = 8,000$ nodes at the next level up

  $8,000/250 = 32$ nodes at the next level up

  1 root node to address the 32 nodes.

  4 levels total, and $1 + 32 + 8,000 + 2,000,000$ nodes total.

- What is the size of the smallest B-tree (in terms of the total number of nodes) that this relation can fit in?

  In this case, we will fill all the nodes as much as we can, i.e. store 500 entries per node.

  At the leaf level, we have $500 * 10^6/500 = 10^6$ nodes.

  At the next level up, we have $10^6/500 = 2,000$ nodes.

  At the next level up, we have $2,000/500 = 4$ nodes.

  1 root node to address the 4 nodes.

- What is the size of the largest B-tree (in terms of the total number of nodes) that this relation can occupy?

  The largest B-tree will have 250 entries per node, so this is equivalent to case 1 above.

Question 4. Suppose you are given a B-tree index for attributes (A,B,C) of R. Each node of the B-tree contains 1,000 entries (except for the root) and the B-tree has height 3 with

a total of 5,000 nodes at the leaf level. What is the cost of answering the following queries by using this index? For each query, list (A) total number of index nodes that need to be scanned, (B) total number of tuples of R that need to be read from R.

- $A = 5$ and $B = 5$ and $C = 5$

  (A) We are going to scan the index for the whole condition. Total number of tuples to be found: $10^5 * 10^3 * 5 * 10^3/10^{12} < 1$, so we expect 0 or 1 tuples. Finding these tuples would cost 1 page from each level of the index, i.e. 4. (B) As the index could be used to answer the whole query, there is no need to read any tuples from R.

- $A = 5$ and $5 \leq B \leq 10$ and $C = 5$

  (A) We are going to scan the index for the condition $A = 5$ and $5 \leq B \leq 10$ and for each tuple we find, we will check $C = 5$. Total number of tuples to scan: $10,000 * 5,000/10^4 = 5,000$ which fits in 5 leaf nodes. Total cost is 3+5 (i.e. one node from each level above leaf).

  (B) As the index could be used to answer the whole query, there is no need to read any tuples from R.

- $A = 5$ and $B = 5$ and $5 \leq C \leq 10$

  (A) We are going to scan the index for the whole condition, i.e. scan $10,000 * 1,000 * 10,000/10^8 = 1,000$ nodes, which requires $3 + 1,000/1,000 = 4$ disk reads from the index.

  (B) As the index could be used to answer the whole query, there is no need to read any tuples from R.

- $5 \leq A \leq 10$ and $B = 5$ and $C = 5$

  (A) We are going to scan the index for the condition $5 \leq A \leq 10$ and for each tuple we find, we are going to check the remaining conditions. Total number of disk nodes to scan $3 + 20,000/1,000 = 23$.

  (B) As the index could be used to answer the whole query, there is no need to read any tuples from R.

- $A = 5$ and $5 \leq D \leq 10$

  (A) We are going to scan the index for $A = 5$ at a cost of $3 + 10,000/1,000 = 13$.

  (B) We need to read each tuple of R we find in (A) from disk to check the D condition. In the worst case, each tuple could be on a different disk page (we are not given any information about how many disk pages R is stored in). As a result, total cost is $10,000$ additional disk pages.

- $5 \leq A \leq 10$ and $5 \leq D \leq 10$

  (A) We are going to scan the index for $5 \leq A \leq 10$ at a cost of $3 + 20,000/1,000 = 23$.

  (B) We need to read each tuple of R we find in (A) from disk to check the D condition. In the worst case, each tuple could be on a different disk page (we are not given any information about how many disk pages R is stored in). As a result, total cost is $20,000$ additional disk pages.