CS5800: Algorithms — Virgil Pavlu

Homework 8

Name:
Collaborators:

Instructions:

- Make sure to put your name on the first page. If you are using the LaTeX template we provided, then you can make sure it appears by filling in the `yourname` command.
- Please review the grading policy outlined in the course information page.
- You must also write down with whom you worked on the assignment. If this changes from problem to problem, then you should write down this information separately with each problem.
- Problem numbers (like Exercise 3.1-1) are corresponding to CLRS 3^{rd} edition. While the 2^{nd} edition has similar problems with similar numbers, the actual exercises and their solutions are different, so make sure you are using the 3^{rd} edition.

1. (50 points)

Implement a hash for text. Given a string as input, construct a hash with words as keys, and word counts as values. Your implementation should include:

- a hash function that has good properties for text
- storage and collision management using linked lists
- operations: insert(key,value), delete(key), increase(key), find(key), list-all-keys

Output the list of words together with their counts on an output file. For this problem,you cannot use built-in-language data structures that can index by strings (like hashtables). Use a language that easily implements linked lists, like C/C++.
You can test your code on "Alice in Wonderland" by Lewis Carroll, at link.
The test file used by TA will probably be shorter.
Try these three values for $m = MAXHASH : 30, 300, 1000$. For each of these m values, produce a histogram over the lengths of collision lists. You can also calculate variance of these lengths.
If your hash is close to uniform in collisions, you should get variance close to zero, and almost all list-lengths around $\alpha = n/m$.
If your hash has long lists, we want to know how many and how long, for example print the lengths of the longest 10% of the lists.
(Extra Credit) Find a way to record not only word counts, but also the positions in text. For each word, besides the count value, build a linked list with positions in the given text. Output this list together with the count.

Solution:

2. (50 points)

Implement a red-black tree, including binary-search-tree operations *sort, search, min, max, successor, predecessor* and specific red-black procedures *rotation, insert, delete.* The *delete* implementation is **Extra Credit** (but highly recommended).

Your code should take the input array of numbers from a file and build a red-black tree with this input by sequence of "inserts". Then interactively ask the user for an operational command like "insert x" or "sort" or "search x" etc, on each of which your code rearranges the tree and if needed produces an output. After each operation also print out the height of the tree.

You can use any mechanism to implement the tree, for example with pointers and struct objects in C++, or with arrays of indices that represent links between parent and children. You cannot use any tree built-in structure in any language.

Solution:

3. Implement Skiplists 50 points

Study the skiplist data structure and operations. They are used for sorting values, but in a datastructure more effficient than lists or arrays, and more guaranteed than binary search trees. Review Slides skiplists.pdf and Visualizer. The demo will be a sequence of operations (asked by

TA) such as for example insert 20, insert 40, insert 10, insert 20, insert 5, insert 80, delete 20, insert 100, insert 20, insert 30, delete 5, insert 50, lookup 80, etc

Solution:

4. (50 points)

Implement binomial heaps as described in class and in the book. You should use links (pointers) to implement the structure as shown in the figure **??**. Your implementation should include the operations: Make-heap, Insert, Minimum, Extract- Min, Union, Decrease-Key, Delete.

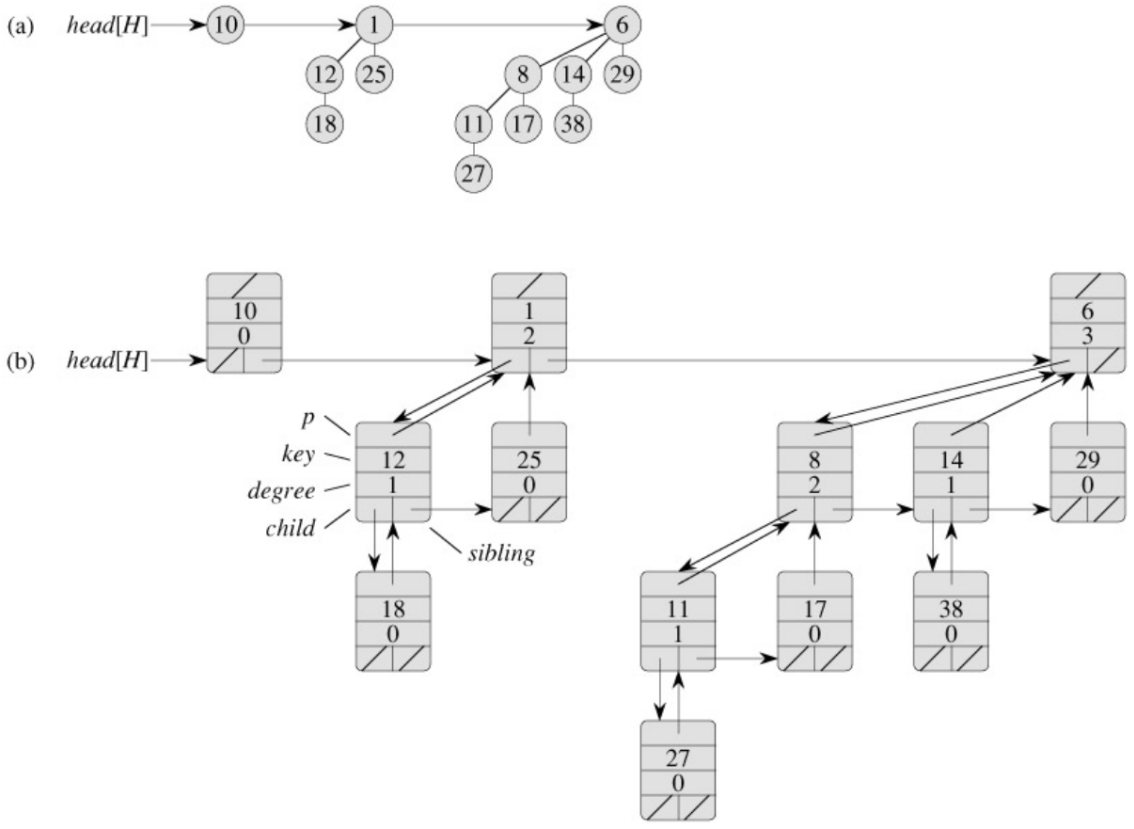

Figure 1: binomial heaps

Make sure to preserve the characteristics of binomial heaps at all times: (1) each component should be a binomial tree with children-keys bigger than the parent-key; (2) the binomial trees should be in order of size from left to right. Test your code with several arrays of randomly generated integers (keys).

Solution:

5. 16.3-3

Consider an ordinary binary min-heap data structure supporting the instructions INSERT and EXTRACT-MIN that, when there are n items in the heap, implements each operation in $O(\log n)$ worst-case time. Give a potential function Φ such that the amortized cost of INSERT is $O(\log n)$ and

the amortized cost of EXTRACT-MIN is $O(1)$, and show that your potential function yields these amortized time bounds. Note that in the analysis, n is the number of items currently in the heap, and you do not know a bound on the maximum number of items that can ever be stored in the heap.

Solution:

1. Basic Definitions

Let's understand what variables we're working with:

- D_i = the heap after ith operation
- n_i = number of elements in D_i
- k = constant for operation time bound
- Each INSERT or EXTRACT-MIN takes at most $k \ln n$ time
- where $n = \max(n_{i-1}, n_i)$

Note: We're using natural log because it will make our calculations cleaner (we'll see why later!)

2. The Potential Function

First, let's define our potential function:

$$\Phi(D_i) = \begin{cases} 0 & \text{if } n_i = 0, \\ k n_i \ln n_i & \text{if } n_i > 0. \end{cases}$$

Key properties:

- Starts at zero: $\Phi(D_0) = 0$ (empty heap)
- Always non-negative: $\Phi(D_i) \geq 0$

3. Proving a Useful Inequality

Before proceeding, we need to prove that for $n \geq 2$: $n \ln \frac{n}{n-1} \leq 2$

Let's break this down step by step:

$$\begin{aligned} n \ln \frac{n}{n-1} &= n \ln \left(1 + \frac{1}{n-1}\right) \\ &= \ln \left(1 + \frac{1}{n-1}\right)^n \\ &\leq \ln \left(e^{\frac{1}{n-1}}\right)^n && \text{(using } 1 + x \leq e^x\text{)} \\ &= \ln e^{\frac{n}{n-1}} \\ &= \frac{n}{n-1} \\ &\leq 2 \end{aligned}$$

4. Analysis of INSERT Operation

4.1 Inserting into Empty Heap

When inserting into empty heap:

- $n_i = 1$
- $n_{i-1} = 0$

Amortized cost calculation:

$$\begin{aligned}\hat{c_i} &= c_i + \Phi(D_i) - \Phi(D_{i-1}) \\ &\leq k \ln 1 + k \cdot 1 \ln 1 - 0 \\ &= 0\end{aligned}$$

4.2 Inserting into Nonempty Heap

When inserting into nonempty heap ($n_i = n_{i-1} + 1 \geq 2$):

$$\begin{aligned}\hat{c_i} &= c_i + \Phi(D_i) - \Phi(D_{i-1}) \\ &\leq k \ln n_i + k n_i \ln n_i - k n_{i-1} \ln n_{i-1} \\ &= k \ln n_i + k n_i \ln n_i - k(n_i - 1) \ln(n_i - 1) \\ &= k \ln n_i + k n_i \ln n_i - k n_i \ln(n_i - 1) + k \ln(n_i - 1) \\ &< 2k \ln n_i + k n_i \ln \frac{n_i}{n_i - 1} \\ &\leq 2k \ln n_i + 2k && \text{(since } n_i \geq 2\text{)} \\ &= O(\lg n_i)\end{aligned}$$

5. Analysis of EXTRACT-MIN Operation

5.1 Extracting Only Item

When extracting the only item:

- $n_i = 0$
- $n_{i-1} = 1$

Amortized cost:

$$\begin{aligned}\hat{c_i} &= c_i + \Phi(D_i) - \Phi(D_{i-1}) \\ &\leq k \ln 1 + 0 - k \cdot 1 \ln 1 \\ &= 0\end{aligned}$$

5.2 Extracting from Heap with Multiple Items

When $n_i = n_{i-1} - 1$ and $n_{i-1} \geq 2$:

$$\begin{aligned}
\hat{c_i} &= c_i + \Phi(D_i) - \Phi(D_{i-1}) \\
&\leq k \ln n_{i-1} + k n_i \ln n_i - k n_{i-1} \ln n_{i-1} \\
&= k \ln n_{i-1} + k(n_{i-1} - 1) \ln(n_{i-1} - 1) - k n_{i-1} \ln n_{i-1} \\
&= k \ln n_{i-1} + k n_{i-1} \ln(n_{i-1} - 1) - k \ln(n_{i-1} - 1) - k n_{i-1} \ln n_{i-1} \\
&= k \ln \frac{n_{i-1}}{n_{i-1} - 1} + k n_{i-1} \ln \frac{n_{i-1} - 1}{n_{i-1}} \\
&< k \ln \frac{n_{i-1}}{n_{i-1} - 1} + k n_{i-1} \ln 1 \\
&= k \ln \frac{n_{i-1}}{n_{i-1} - 1} \\
&\leq k \ln 2 && \text{(since } n_{i-1} \geq 2\text{)} \\
&= O(1)
\end{aligned}$$

6. An Alternative Potential Function

Let's look at another way to define the potential function:

For each node x in heap:

- Let $d_i(x)$ = depth of node x in D_i
- Define:

$$\begin{aligned}
\Phi(D_i) &= \sum_{x \in D_i} k(d_i(x) + 1) \\
&= k\left(n_i + \sum_{x \in D_i} d_i(x)\right)
\end{aligned}$$

Properties of this new function:

- Initially $\Phi(D_0) = 0$ (empty set sum)
- Always $\Phi(D_i) \geq 0$
- After INSERT: changes by $k(1 + \lfloor \lg n_i \rfloor)$
  - Amortized cost: $O(\lg n_i) + O(\lg n_i) = O(\lg n)$
- After EXTRACT-MIN: decreases by $k(1 + \lfloor \lg n_{i-1} \rfloor)$
  - Amortized cost: $k \lg n_{i-1} - k(1 + \lfloor \lg n_{i-1} \rfloor) = O(1)$