

**Advanced Algorithm**

**DR Moeini**

**Parsa Mohammadpour**

**Hesam Mumivand Fard**

**Mohammad Rahbarimoghadam**

**Saeid Terik**

**Second assignment**

1. Prove that it doesn’t matter which node you start searching from in the binary search tree with the height of h, k times of calling TREE\_SUCCESSOR, will be done in O (h+k).
2. Assume that instead of having x.p (nodes parent) at each node that refers to the node parent, each node maintains the variable x.succ that refers to the node successor. By writing pseudocode of SEARCH, INSERT and DELETE in binary search tree show how they change. These actions must be done in O (h), where h is the height of the tree. (help: try to write the pseudocode that returns each nodes parent)
3. Consider that we have alphabet {a, b, c, d, e}. Assume that we have the following frequencies for this alphabet in a text:

What is the number of expected bits for Huffman encoding of a text with 1000 character?

First, we briefly explain Huffman encoding. Then we show the result of the algorithm for this example.

Huffman algorithm: Huffman algorithm is a greedy algorithm designed for encoding and decoding efficiently. This algorithm encodes each alphabet member such that no two different characters are prefix of each other and also tries to minimum the number of total bits that are used for encoding the whole text. As we said, this algorithm is a greedy algorithm. But why? We will see about that in the algorithm description. This goal, having minimum number of the bits in the encoded document, we try to assign the minimum length of the encoded bit-string[[1]](#footnote-1) (string containing zeros and ones) to the alphabet that has the minimum frequency.

Huffman algorithm description: In this algorithm, we want to make a tree from these alphabets, a tree like BST[[2]](#footnote-2), but with differences, and keep alphabet characters in its leaves and then by applying a simple rule we assign a bit-string to each of these characters. Our rule for assigning the bit-string to characters is to start from the root, if you went on the left child, add zero to the bit-string (which is empty string at the beginning) and by going on the right child, add one to the bit-string. Do this action until you get to the character that you want.

So, in the algorithm, first we convert each character with its frequency to a BST node. These nodes contain the character and the frequency of that character. For this purpose, we use a min-heap[[3]](#footnote-3) (minimum priority queue) to store these nodes and the comparison function used for this heap, is by comparing each node frequency. In other meaning, we store a min-heap of these nodes by comparing their frequencies. Then at each step of the algorithm, we get first element of the heap by getting calling the *pop[[4]](#footnote-4)* function of the heap, we get the element of the heap with least frequency and then by doing this action again, we get the element with second least frequency from the heap. Then we merge these two nodes together. We define merging two node, like the follow:

Merging two nodes: Generate a new node and set the frequency of this new node to the summation of these two nodes. And then assign the node with the smaller frequency to the left child of this new node and assign the node with higher frequency to the right child of this new node. Then put this new node to the heap.

So, at each step, we remove two nodes from heap, and then add another node in that heap hence we have n-1 step totally. All these actions are done in O (1) except for heap actions. Heap actions like adding element and removing elements are done in O (lg n). So, the overall time complexity of this algorithm by using heap, is O (n lg n). We also have to generate a heap from these nodes at the beginning. This action is done in O (n). So, the overall time complexity is O (n + n lg n) which is same as O (n lg n).

Now let’s move on to our example. By the algorithm, first we make nodes with frequency of each character and setting that character value in that node. Then we have to make a heap from these nodes. But for simplification, we just use the sorted array. Because their difference was in the time complexity not in the functionality of the algorithm. So, our sorted array is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Node | a | b | c | d | e |
| Freq | 0.32 | 0.25 | 0.2 | 0.18 | 0.05 |

So as the algorithm requires, we have to get two nodes with minimum frequencies from this array, which are d and e, and then merge them together with each other. So, the merges node is as follows:

Merging e and d, will result in a new node f which has frequency of 0.05+0.18 = 0.23. SO, after inserting this new node to the array, the array is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Node | a | b | f  e d | c |
| Freq | 0.32 | 0.25 | 0.23 | 0.2 |

Now, if we apply this algorithm for the next step, we will have a new node g, which has the frequency of 0.23+0.2 which is equal to 0.43. So, the sorted array is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Node | g  c f  e d | a | b |
| Freq | 0.43 | 0.32 | 0.25 |

Now we move forward to another step for the algorithm. So, we now make a new node from merging nodes a and b and name that new node to h. The frequency of the h is 0.25+0.32 which is equal to 0.57. So our sorted array will be as follows:

|  |  |  |
| --- | --- | --- |
| Node | h  a b | g  c f  e d |
| Freq | 0.57 | 0.43 |

Now, for the last step we get these last two remaining nodes and merge them together. So, now get these two nodes and merge them together and make the final node, the root of the tree, and then we will explain the final tree and assign the bit-string to each leaf (each alphabet character) and then compute the final answer. So, first let’s build final tree. Our tree will be as follows:

|  |  |
| --- | --- |
| Node | i  h g  a b c f  e d |
| Freq | 1 |

So, now we will have to assign the bit-string to each character. So, we apply the bit-string assignment algorithm, which was starting from the root with an empty string, and then for each time that go to the left child, we add a zero to the bit-string and each time we go to the right child, we add one to the bit-string. Then bit-string of each character is as follows:

|  |  |
| --- | --- |
| character | bit-string |
| a | 00 |
| b | 01 |
| c | 10 |
| d | 111 |
| e | 110 |

Now, for the last step, we have to find expected number of repetitions in the document and multiply that with number of required bit-string that is assigned to each character. Then we will compute some extra computation to show that this algorithm, Huffman algorithm, is more efficient.

So, due to the provided frequency of the characters, we will compute the expected number of bit used to encode each part of the document that is related each character. These computation for each character is as follows:

* character a:
* character b:
* character c:
* character d:
* character e:

So, the expected value of total number of used bits, is equal to 640+500+400+540+150 = 2230 bits.

But, if we wanted to use fixed length encoding like ASCII nut with length of 3 bits for all characters, we would have the exact value of 10003 = 3000 bits. So, obviously, by using Huffman algorithm, we have improved the expected number of bits in the document.

1. bit-string (binary string): the string containing only zeros and ones [↑](#footnote-ref-1)
2. BST: binary search tree [↑](#footnote-ref-2)
3. using heap is not mandatory, we also could use array and sort that array at each step, but it would have more time complexity. [↑](#footnote-ref-3)
4. pop function of the min-heap, return the minimum element of the heap and then remove it. Then by setting the last element to the first element and by applying some functions like *bubble-down* and *bubble-up*, it re-generates the heap such that the first element is the minimum element of the heap. [↑](#footnote-ref-4)