Internal path length of the binary representation of heap-ordered trees [☆]

Wen-Chin Chen [*], Wen-Chun Ni

Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, ROC

Abstract

We show that the average internal path length of the binary representation of a random n-node heap-ordered tree is asymptotically $n\sqrt{\pi n} - 2n + O(\sqrt{n})$.

Key words: Algorithms; Combinatorial enumeration; Data structures; Generating functions; Heap-ordered trees; Internal path length; Probability

1. Introduction

This paper studies the *heap* (or *priority queue*). Heap serves as the underlying data structure for the efficient implementation of many combinatorial algorithms such as shortest paths, minimum spanning trees, and network flows. A *heap-ordered tree* (HOT) is defined recursively to be a *rooted labeled* tree whose root contains the minimum key (label) and whose subtrees are themselves HOTs. An alternative definition is that each node in a HOT has key value no greater than its child nodes. Throughout, we assume that all the key values of an $(n + 1)$-node HOT are distinctly chosen (labeled) from the set

$\{0, 1, \ldots, n\}$. Obviously the root of an HOT is always labeled 0.

In computer applications, a multiway tree can be coded as a binary tree representation as follows: the first child and the right sibling of a node in a tree becomes the left-child and the right-child of the node in the corresponding binary representation, respectively. For example, Fig. 1(b) gives the binary tree representation of the HOT in Fig. 1(a). Since the root of the binary representation of a HOT has a null right subtree, it is convenient to consider the variant of HOT obtained by deleting its root node. Such a variant is called *heap-ordered forest* (HOF) because it can be viewed as a collection of HOTs. Note that the heap property that each node has minimum key value in the subtree rooted at this node should be modified slightly in the binary representation. In the binary representation, the heap property only requires that each node has key value no greater than the

[*] Corresponding author.

keys in its left-subtree. The key values in the right-subtree are irrelevant to the node. Since we assume previously that an $(n + 1)$-node HOT is labeled from the set $\{0, 1, \ldots, n\}$, an n-node HOF is labeled from the set $\{1, 2, \ldots, n\}$.

The *internal path length* (IPL) of a HOT is defined to be the sum of the *path lengths* of all the nodes in the binary representation of the tree, where the path length of a node v is the number of edges in the path from the root to v. For example, the path length of node **2** in the binary tree in Fig. 1(b) is 4 and the IPL of the HOT is 25. IPL is an important quantity for analyzing the average time complexity of the *search* and *find _min* operations of heap.

This paper derives the average IPL of a random n-node HOT using the techniques of recurrence and generating functions. It shows that the average IPL of a random n-node HOT is asymptotically $n\sqrt{\pi n} - 2n + O(\sqrt{n})$.

2. Derivation of internal path length

Let a_n and b_n be the total number of n-node HOTs and n-node HOFs, respectively. It was shown in [1] that a_n, b_n, and their exponential generating functions $A(x) = \sum_{n \geqslant 0} a_n x^n / n!$ and $B(x) = \sum_{n \geqslant 0} b_n x^n / n!$ satisfy the following proposition.

Proposition 1.

$$a_n = \frac{(n-1)!}{2^{n-1}} \binom{2n-2}{n-1},$$

$$b_n = \frac{n!}{2^n} \binom{2n}{n},$$

$$A(x) = 1 - \sqrt{1 - 2x},$$

$$B(x) = 1/\sqrt{1 - 2x} = (1 - A(x))^{-1} = A'(x).$$

To derive the average IPL of an n-node HOT, we let $a_{n,p}$ and $b_{n,p}$ denote respectively the total number of n-node HOTs and HOFs whose internal path length is p. The double generating functions of $a_{n,p}$ and $b_{n,p}$,

$$A(x, y) = \sum_{n,p} a_{n,p} y^p x^n / n!,$$

$$B(x, y) = \sum_{n,p} b_{n,p} y^p x^n / n!,$$

will be the main vehicle for our derivation of the average IPL. Note that it is easy to see that $A(x, 1) = A(x)$ and $B(x, 1) = B(x)$. If we let $e_n = \sum_{p \geqslant 0} p a_{n,p}$, then the average IPL of the n-node HOTs is e_n / a_n. Thus, it suffices to compute e_n and its exponential generating function,

$$E(x) = \sum_n e_n x^n / n! = A_y(x, 1). \tag{1}$$

Since the binary representation of an n-node HOF F is obtained by deleting the root node of the binary representation of the corresponding
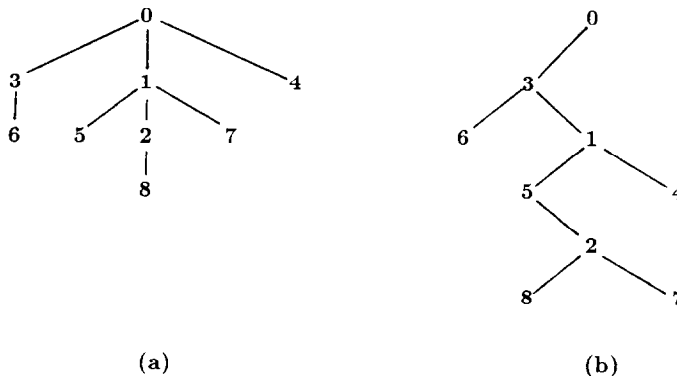

(a)   (b)

Fig. 1. (a) A HOT; and (b) its binary representation.

$(n + 1)$-node HOT T, the path length of the nodes in F is 1 less than their path length in T. We thus have $a_{n+1,p} = b_{n,p-n}$, and the two generating functions $A(x, y)$ and $B(x, y)$ can be related as

$$A_x(x, y) = \sum_{n,p} a_{n+1,p} y^p x^n / n!$$

$$= \sum_{n,p} b_{n,p-n} y^{p-n} (xy)^n / n!$$

$$= B(xy, y). \tag{2}$$

We use the notations $A_x(x, y) = \partial A(x, y)/\partial x$ and $A_y(x, y) = \partial A(x, y)/\partial y$ for convenience.

Our derivation of average IPL is based mainly on the following recurrence of $b_{n,p}$ ($n \geqslant 1$):

$$b_{n,p} = \sum_{1 \leqslant i \leqslant n} \sum_{\substack{0 \leqslant k \leqslant n-i \\ 0 \leqslant r \leqslant p-n+1}} \binom{n-i}{k}$$

$$\times b_{k,r} b_{n-k-1,p-r-n+1}.$$

with initial condition $b_{0,p} = \delta_p = 0$, where the *Kronecker delta function* function δ_R denotes 1 if relation R is true and 0 otherwise. The recurrence holds for the following reason. Let T be the left-most HOT of an n-node HOF F. If the root node of T is labeled i and T contains $(k + 1)$ nodes, then there are $\binom{n-i}{k}$ ways to choose k nodes into T. Also, if the IPLs of F and T are p and r, respectively, then the IPL of the HOF obtained by removing T from F is $p - r - n + 1$.

Using the techniques developed in [3], the above recurrence of $b_{n,p}$ can be simplified as

$$b_{n,p} = \sum_{\substack{1 \leqslant k \leqslant n \\ 0 \leqslant r \leqslant p-n+1}} \binom{n}{k} b_{k-1,r} b_{n-k,p-r-n+1}.$$

The initial condition of $b_{n,p}$ can be absorbed into the recurrence by adding a delta term:

$$b_{n,p} = \delta_{n=p=0}$$

$$+ \sum_{\substack{1 \leqslant k \leqslant n \\ 0 \leqslant r \leqslant p-n+1}} \binom{n}{k} b_{k-1,r} b_{n-k,p-r-n+1}.$$

This recurrence leads to the equation of $B(x, y)$:

$$B(xy, y) \int_0^x B(ty, y) \, \mathrm{d}t$$

$$= \sum_{n,p} b_{n,p} \frac{y^p (xy)^n}{n!} \sum_{n,p} b_{n,p} \frac{y^{p-1}(xy)^{n+1}}{(n+1)!}$$

$$= \sum_{n,k,p,r} \binom{n}{k} b_{k-1,r} b_{n-k,p-r} \frac{x^n y^{p+n-1}}{n!}$$

$$= \sum_{n,k,p,r} \binom{n}{k} b_{k-1,r} b_{n-k,p-r-n+1} \frac{x^n y^p}{n!}$$

$$= B(x, y) - 1.$$

Plugging (2) into the above equation, we have

$$A_x(x, y) A(x, y) = A_x(x/y, y) - 1.$$

It follows then that

$$\tfrac{1}{2} A^2(x, y) = y A(x/y, y) - x.$$

Namely,

$$\tfrac{1}{2} A^2(xy, y) - y A(x, y) + xy = 0. \tag{3}$$

The closed form of $A(x, y)$ is not easy to find. However, to find the average internal path length, we only have to compute $E(x) = A_y(x, 1)$. This can be done by differentiating (3) with respect to y and then setting $y = 1$ to have

$$A(x, 1)(x A_x(x, 1) + A_y(x, 1))$$

$$- A(x, 1) - A_y(x, 1) + x = 0. \tag{4}$$

Since

$A(x, 1) = A(x) = 1 - 1/B(x)$
    from Proposition 1,
$A_x(x, 1) = B(x, 1) = B(x)$ from (2),
$A_y(x, 1) = E(x)$ from (1),

Eq. (4) gives

$$E(x) = x B^2(x) - B(x) + 1$$

$$= \sum_{n \geqslant 1} (n! 2^{n-1} - b_n) x^n / n!.$$

Thus $e_n = n!2^{n-1} - b_n$ and the average internal path length of the n-node HOTs ($n > 0$) equals to

$$\frac{e_n}{a_n} = \frac{4^{n-1}n}{\binom{2n-2}{n-1}} - (2n-1)$$

$$= \sqrt{\pi}\, n^{3/2} - 2n - \tfrac{3}{8}\sqrt{\pi}\, n^{1/2}$$

$$+ 1 - \tfrac{7}{128}\sqrt{\pi}\, n^{-1/2} + \mathrm{O}(n^{-3/2}).$$

We thus have:

Theorem 2. *The average internal path length of the binary representation of the n-node HOTs is*

$$\sqrt{\pi}\, n^{3/2} - 2n - \tfrac{3}{8}\sqrt{\pi}\, n^{1/2}$$

$$+ 1 - \tfrac{7}{128}\sqrt{\pi}\, n^{-1/2} + \mathrm{O}(n^{-3/2}).$$

3. Conclusions

It was shown in [1,2,4] that the average altitude (or level) of the nodes in a random n-node heap-ordered tree is asymptotic to 0.5 ln n. Our result in this paper is rather surprising because it contradicts the general belief that the binary representation of the heap-ordered trees has the average height only "twice" that of the general trees.

References

[1] F. Bergeron, P. Flajolet and B. Salvy, Varieties of Increasing Trees, in: J.-C. Raoult, ed., *Proc. CAAP 1992, 17th Coll. on Trees in Algebra and Programming*, Rennes, France, Lecture Notes in Computer Science (Springer, Berlin, 1992) 24–48.

[2] W.C. Chen and W.C. Ni, On the average altitude of heap-ordered trees, *Internat. J. Foundation Comput. Sci.* to appear.

[3] R.L. Graham, D.E. Knuth and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science* (Addison-Wesley, Reading, MA, 1989).

[4] H.M. Mahmoud, Distances in random plane-oriented recursive trees, *J. Comput. Appl. Math.* **41** (1992) 237–245.