The Expected Distribution of Degrees in Random Binary Search Trees

H. M. MAHMOUD
The George Washington University, Department of Statistics, Washington DC 20052, USA

Let V_i be the set of vertices of degree i, i = 1,2,3, in a random binary search tree. We prove that $E(|V_i|) = n/3 + 0(1)$, for i = 1,2,3. This result tells us that the expected tree shape does not contain very long path subgraphs; thus in a sense the expected shape tends to be balanced.

Received February 1984

1. INTRODUCTION

In many practical situations we construct binary search trees (BSTs) from input sequences on-line (as they progress in time). It is a fairly realistic assumption that all orderings of the input are equally likely.[1,2]

We may consider, without loss of generality, that the input stream is a permutation of the integers 1,2,...,n. (For definitions, properties and construction algorithms of BSTs the reader is referred to any textbook on data structures, for example Refs 1, 3 or 4.)

In this note, when we say random binary search tree we mean the randomness induced on the space of trees by considering all $n!$ permutations of the input as equally probable.

2. THE MAIN RESULT AND ITS INTERPRETATION

We investigate the expected distribution of node degrees in random binary search trees, and we show that the sizes of the sets of nodes of degree i, $i = 1,2,3$, are asymptotically equivalent to $n/3$, n being the number of nodes in the tree. This result tells us that the 'expected' tree shape does not contain very long path subgraphs, because a tree with such a path would have a lot of nodes of degree two and a small number or vertices of degree one or three. Thus, in a sense, the expected shape of a binary search tree tends to be balanced.

3. THE EXPECTED NUMBER OF LEAVES

The leaves of an unextended binary tree are the terminal nodes encountered in any path starting at the root. Using techniques similar to those of Hibbard,[5] we first find the expected number of leaves in BSTs. The number of leaves is obviously related to the number of nodes of degree one in the tree depending on whether the root has degree one, too. Similar work was done in[6] for the class of recursive trees.

Definition

Let e_n be the random number of leaves in a random (unextended) binary search tree, and let $E(e_n)$ be the expected value of this random variable. We denote $E(e_n)$ *as* E_n for brevity.

Lemma 1

$$E_n = \left(\frac{n+1}{3}\right) \quad \text{for} \quad n > 2 \quad \text{and} \quad E_2 = 1,\ E_1 = 1.$$

Proof

When $n = 1$ the tree is a single node which is also a leaf, thus $E_1 = 1$. When $n = 2$ we have two permutations giving the trees of Fig. 1; E_2 is clearly equal to one.

Figure 1. The trees corresponding to the permutations of the set {1,2}

For $n > 2$, let $t_n^{(l)}$, $t_n^{(r)}$ be the left and right subtrees respectively. Denote by $U_n^{(s)}$ the number of nodes of $t_n^{(s)}$ ($s \in \{l, r\}$). Clearly

$$U_n^{(l)} + U_n^{(r)} = n - 1.$$

Also, conditioned on the event $(U_n^{(l)} = i-1)$, $t_n^{(l)}$ and $t_n^{(r)}$ will be distributed as t_{i-1} and t_{n-i}, respectively, $(1 \leqslant i \leqslant n)$. This results from the insertion algorithm and the probability measure on permutations. (The event $(U_n^{(l)} = i-1)$ is equivalent to the event that i appears first in the permutation.) Hence,

$$\begin{aligned} E(e_n \mid U_n^{(l)} = i-1) &= E(e_{i-1} + e_{n-i} \mid U_n^{(l)} = i-1) \\ &= E(e_{i-1}) + E(e_{n-i}) \\ &= E_{i-1} + E_{n-i}. \end{aligned}$$

Our basic assumption that all n-permutations are equally likely implies that $P(U_n^{(l)} = i-1) = \frac{1}{n}$, $1 \leqslant i \leqslant n$. Therefore, the unconditional expectation of e_n is

$$E_n = \frac{1}{n} \sum_{i=1}^{n} (E_{i-1} + E_{n-i})$$

or

$$nE_n = 2 \sum_{i=1}^{n} E_{i-1}. \tag{3.1}$$

To solve this recurrence, compare a version of (3.1) with $n+1$ replacing n with (3.1) to get

$$E_n = \frac{(n+1)}{n} E_{n-1}.$$

The recurrence is now in a form suitable for iterative substitution:

$$\begin{aligned} E_n &= \frac{n+1}{n} E_{n-1} \\ &= \frac{n+1}{n} \frac{n}{n-1} E_{n-2} \\ &\ \ \vdots \\ &= \frac{n+1}{n} \frac{n}{n-1} \frac{n-1}{n-2} \cdots \frac{5}{4} \frac{4}{3} E_2 \\ &= \frac{n+1}{3} \quad \square \end{aligned}$$

4. THE SHAPE OF A BINARY SEARCH TREE

Consider the undirected graph underlying the directed tree. The leaves are nodes of degree one, and non-terminal nodes may have degrees two or three as in Fig. 2. The root may have degree one, however, since it has no 'parent' in the tree.

Nodes of degree two Nodes of degree three

Figure 2. Two types of non-terminal nodes

Definition

Let V_i be the set of vertices of degree i, $i = 1,2,3$, and let v_i be the expected number of vertices of degree i, $i = 1,2,3$.

Theorem

$$v_i = \frac{n}{3} + 0(1), \quad i = 1,2,3.$$

Proof

For any tree we have

$$|V_1| = \begin{cases} e_n + 1 & \text{if the root has degree 1,} \\ e_n & \text{if the root has degree 2.} \end{cases}$$

Thus the unconditional expectation $E(|V_1|) = v_1$ is given by

$$\begin{aligned} &E(e_n+1)*P(\text{the root has degree 1}) \\ &+E(e_n)*P(\text{the root has degree 2}) \\ &= E(e_n)+P(\text{the root has degree 1}). \end{aligned}$$

Using the notations of the previous lemma

$$\begin{aligned} P(\text{the root has degree 1}) &= P(U_n^{(l)} = 0) \\ &+ P(U_n^{(r)} = 0) = \frac{2}{n}. \end{aligned}$$

Thus

$$v_1 = E_n + \frac{2}{n} = \frac{n+1}{3} + \frac{2}{n}. \tag{4.1}$$

It is well known (see Ref. 7, for example) that for any graph $G = (V, \mathscr{E})$:

$$\sum_{v \in V} d(v) = 2|\mathscr{E}|. \tag{4.2}$$

Here, V and $\mathscr{E}$ are the sets of vertices and edges and $d(v)$ is the degree of vertex v in G. Also, for any tree $T = (V, \mathscr{E})$ (see Ref. 7), the cardinalities of V and $\mathscr{E}$ are related by

$$|\mathscr{E}| = |V| - 1. \tag{4.3}$$

Thus (4.2) and (4.3) for a binary tree yield

$$|V_1| + 2|V_2| + 3|V_3| = \sum_{v \in V} d(v) = 2(n-1). \tag{4.4}$$

The fact that V is the disjoint union of V_1, V_2, V_3 yields

$$|V_1| + |V_2| + |V_3| = |V| = n. \tag{4.5}$$

Taking expectations of (4.5) and (4.4) together with the notation introduced before the lemma for the expected values, we get

$$v_1 + v_2 + v_3 = n, \tag{4.6}$$

$$v_1 + 2v_2 + 3v_3 = 2n - 2. \tag{4.7}$$

The linear system (4.1), (4.6) and (4.7) has the solution

$$v_1 = \frac{n+1}{3} + \frac{2}{n},$$

$$v_2 = \frac{n+4}{3} - \frac{4}{n},$$

$$v_3 = \frac{n-5}{3} + \frac{2}{n};$$

and the theorem follows □

Acknowledgements

The author is grateful to an anonymous referee whose comments helped him improve notations and remove some ambiguities in an earlier manuscript.

REFERENCES

1. D. E. Knuth. *The Art of Computer Programming*, vol. 3. Addison–Wesley, Reading, Massachusetts. (1973).
2. H. Mahmoud. A probabilistic analysis of a class of random trees, *Ph.D. thesis*, the Ohio State University, Columbus, Ohio. (1983).
3. E. Horowitz and S. Sahni. *Fundamentals of Data Structures in PASCAL*. Computer Science Press, Potomac, Maryland. (1984).
4. A. Aho, J. Hopcroft and J. Ullman. *Data Structures and Algorithms*. Addison-Wesley, Reading, Massachusetts. (1983).
5. T. N. Hibbard. Some combinatorial properties of certain trees with applications to searching and sorting. *Journal of the ACM* **9**, 13–28 (1962).
6. M. Dondajewski and J. Szymanski. On the distribution of vertex degrees in a *Strata of a random recursive tree, Bulletin de l'Académie Polonaise des Sciences, Série des Sciences Mathématiques*, vol. xxx, nos. 5–6 (1982).
7. J. Bondy and U. Murty. *Graph Theory with Applications*. North-Holland, New York. (1980).