Efficiency of a Binary Comparison Storage Technique

E. M. PALMER, M. A. RAHIMI, AND R. W. ROBINSON

University of Michigan, Ann Arbor, Michigan

ABSTRACT. The efficiency of an information storage technique based on binary comparisons is analyzed. Generating functions are applied to finding the mean and variance of the number of comparisons needed to retrieve one item from a store of n items. Surprisingly, the variance approaches $7 - \frac{2}{3}\pi^2$ for large n.

KEY WORDS AND PHRASES: searching, binary trees, information retrieval, trees, directory search, generating functions, mean, variance

CR CATEGORIES: 3.74, 5.31, 5.32

Introduction

We analyze the efficiency of an information storage technique for randomly acquired data which is tagged by some naturally ordered set of labels. The array of stored data is enlarged as new items are entered, but once stored no part is ever rearranged. A binary search tree is an abstraction of this array.

The storage structure can be visualized as a binary tree in which successive items are stored on limbs branching to the left or right according to whether their labels are smaller or greater in their linear order. Retrieval involves comparing the wanted label with certain of those which have been stored; the number of these comparisons is a measure of the cost of retrieval.

Windley [3] was the first to systematically study the search trees considered in the present paper. He solved a recurrence relation for the average number of comparisons required to retrieve one item from a store of n. This result was also obtained by Martin and Ness [2], and a parallel result for a slightly different class of search trees was found by Hibbard [1]. In addition, Windley [3] discussed the variance of the number of comparisons. He found a recurrence relation from which he computed the variance for particular values of n, and observed empirically that the variance is small compared to the mean.

In this article we also calculate the mean number of comparisons. We have, however, expressed our recurrence relation in terms of generating functions. Then, on solving a linear differential equation of first order we find the mean number of comparisons to be $1.386 \log_2 n - 2.846 + R_n$, where $R_n < 1$ for $n > 5$ and $\lim_{n\to\infty} R_n = 0$.

The most favorable order of receipt of the data gives an array in which this average is

The work done by E. M. Palmer was supported in part by a grant from the NSF. The work done by R. W. Robinson was supported in part by grant number 73-2502 from the US Air Force Office of Scientific Research.

Authors' present addresses: E. M. Palmer, Department of Mathematics, Michigan State University, East Lansing, MI 48824; M. A. Rahimi, Computer Science Department, Michigan State University, East Lansing, MI 48824; R. W. Robinson, Department of Mathematics, University of Newcastle, Newcastle, NSW 2308, Australia.

$\log_2 n - 2 + R_n'$, where the remainder R_n' asymptotically varies between 0 and 0.1. Thus for many applications it may be uneconomical to rearrange the data once it has been stored.

The computation of the variance of the averages is greatly facilitated by the use of generating functions. Surprisingly, this variance approaches a constant limit of $7 - \frac{2}{3}\pi^2 = .420\cdots$. Thus it is highly unlikely that a randomly obtained data storage array is significantly different in efficiency from the average.

1. *Data Storage*

The data storage technique which we analyze presumes that each item of information is tagged, and that the tags have some natural linear order. For instance, if the data should be medical or other information about individuals, the tag could be the name and date of birth of the person involved. Names are naturally ordered lexicographically as in a telephone directory, with ties to be broken by date of birth.

When an item is stored, say at address A, its tag l_1 is listed first, then two spaces s_1, s_2 for addresses, and then the rest of the information. Every new item is directed in sequence to a series of addresses before it is finally stored. If the item was labeled l_2 and is directed to the address A, we examine the contents of s_1 or s_2 according to whether $l_2 < l_1$ or $l_1 < l_2$. If the space contains an address, the new item is directed to that address. If the space is empty, the new item is stored at some address of opportunity, and this address is entered in the previously empty space.

For example, suppose that our data consists of a short biography of each of the U.S. presidents of the twentieth century to date, stored in the order in which they held office: McKinley Roosevelt, T. Taft Wilson Harding Coolidge Hoover Roosevelt, F. Truman Eisenhower Kennedy Johnson Nixon. Suppose further that the name of each president is used as the tag for his biography, comparison of names to be lexicographic. The first item stored would be the ordered quadruple consisting of the name McKinley, space s_1, space s_2, and the biography. At first empty, space s_1 eventually is occupied by the address of Harding's entry. Space s_2 is taken by T. Roosevelt's entry. The finished configuration is represented as a rooted binary tree in Figure 1, with McKinley as the root. To search the tree for the biography of a particular president, the president's name is used in a series of comparisons as if it were the tag of a new piece of information, until his biography is located. The number of comparisons required before the biography is located is just its distance from the root. Note that in Figure 1 the average distance from the root is 30/13. Evidently the least possible average distance

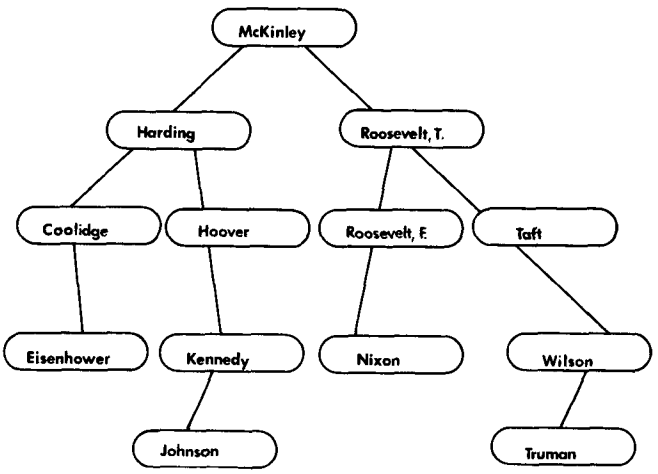

FIG. 1. Data tree for the twentieth century presidents

for any tree corresponding to 13 items is 28/13, while the worst possible (for a linear tree) is 78/13.

The rooted binary tree associated with this list of names depends only on the natural (alphabetic) order of the tags (names) and the order in which they appear. Similarly any sequence $a_1, \cdots, a_n$ of n different positive integers has associated with it a binary tree which depends on the natural order of the integers and the order in which they appear, say from left to right, in the sequence. For example, the sequence 10, 15, 16, 20, 3, 1, 5, 13, 17, 2, 8, 7, 11 results in the tree of Figure 2. Since 10 is first in the sequence, it is at the root; 15, appearing next, is greater than 10 and hence takes the point adjacent to 10 and to the right of 10, etc.

Note that many different sequences can result in the same binary tree. For example, the sequence 10, 15, 13, 3, 16, 5, 1, 11, 8, 20, 2, 7, 17 is also associated with the tree of Figure 2.

2. *Mean Number of Comparisons*

Each sequence of n different items has as a measure of the cost of retrieval, the average number of comparisons required to find one item. Therefore we consider the random variable which assigns to each of the $n!$ equiprobable sequences this average, which can also be interpreted as the mean distance from the root of the points of the corresponding binary tree. The next theorem provides an exact formula for the mean of the distribution under consideration, denoted $d_n/n!n$.

THEOREM 1. *The mean distance from the root of the $n!n$ points in the binary trees determined by all $n!$ sequences of n different positive integers is*

$$d_n/n!n = 2(1 + 1/n)\left(\sum_{k=1}^{n} (1/k)\right) - 4. \tag{1}$$

This result is simplified when expressed in terms of Euler's constant $\gamma = .577\cdots$. By definition it is well known that

$$\sum_{k=1}^{n} (1/k) = \log n + \gamma + O(1/n). \tag{2}$$

COLLOLLARY 1. *The mean number of comparisons required to retrieve one item from a store of n is*

$$d_n/n!n = 1.386 \log_2 n - 2.846 + R_n. \tag{3}$$

In Corollary 1, $1.386\cdots = 2 \log 2$, $2.846\cdots = 2\gamma - 4$, and $R_n = O(1/n)$.

To prove Theorem 1 we can assume without loss of generality that our sequences involve only the numbers from 1 to n. Now we express d_n, the total distance from the root of all $n!n$ points, in terms of d_k with $k < n$:

$$d_n = \sum_{p=1}^{n} \binom{n-1}{p-1}[(n-p)!\, d_{p-1} + (p-1)!\, d_{n-p} + (p-1)!(n-p)!(n-1)]. \tag{4}$$

To see this, we first compute the distances in trees corresponding to sequences $a_1, \cdots, a_n$ with $a_1 = p$, and then sum over $p = 1, \cdots, n$. For a fixed sequence with $a_1 = p$, let

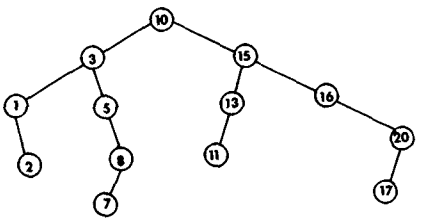

FIG. 2. A binary tree associated with a sequence of numbers

A be the total distance in the tree on $p - 1$ points associated with the subsequence of numbers less than p. Similarly let B be the corresponding distance contributed by the numbers greater than p. Then the sum of the distances in the entire tree is

$$A + B + (n - 1). \tag{5}$$

The term $(n - 1)$ accounts for the extra comparison with a_1 made for $a_2, \cdots, a_n$. Now the right-hand factor of (4) is the result of summing the expression (5) over all $(p - 1)!$ possible orderings of the subsequence of numbers less than p and independently over all $(n - p)!$ orderings of the numbers greater than p. The result of summing A over just the $(p - 1)!$ orderings is d_{p-1}, while the sum of B over the $(n - p)!$ orderings is d_{n-p}. Finally, the factor $\binom{n-1}{p-1}$ is the number of ways that the two subsequences can be interlaced.

A more compact form is obtained by dividing both sides of eq. (4) by $(n - 1)!$ and simplifying with the observation that after summation, the contributions corresponding to A and B are equal:

$$nd_n/n! = n(n - 1) + 2\sum_{k=0}^{n-1}(d_k/k!). \tag{6}$$

This equation suggests the use of the exponential generating function $d(x)$ defined by

$$d(x) = \sum_{n=2}^{\infty} d_n x^n/n!. \tag{7}$$

Note that the formal derivative $d'(x)$ has the left side of (6) as the coefficient of x^{n-1}. The right side can also be expressed in terms of generating functions. We use the simple observation

$$1/(1 - x) = \sum_{k=0}^{\infty} x^k, \tag{8}$$

from which it follows by successive differentiation that

$$1/(1 - x)^2 = \sum_{k=1}^{\infty} kx^{k-1} \tag{9}$$

and

$$2/(1 - x)^3 = \sum_{k=2}^{\infty} k(k - 1)x^{k-2}. \tag{10}$$

Now note that $2\sum_{k=0}^{n-1} d_k/k!$ is the coefficient of x^{n-1} in $2d(x)/(1 - x)$ while from (10) it follows that $n(n - 1)$ is the coefficient of x^{n-1} in $2x/(1 - x)^3$. Thus we find that $d(x)$ is a solution of the differential equation

$$y' = 2y/(1 - x) + 2x/(1 - x)^3. \tag{11}$$

On transferring to the left side those terms involving y and multiplying by the integrating factor $(1 - x)^2$, we have

$$(1 - x)^2y' - 2y(1 - x) = 2x/(1 - x), \tag{12}$$

or

$$(d/dx)(y(1 - x)^2) = 2x/(1 - x). \tag{13}$$

On integrating this equation, the result is

$$y(1 - x)^2 = 2\int [x/(1 - x)]\,dx = 2\int \left(\sum_{k=1}^{\infty} x^k\right)dx = 2\left\{\sum_{k=1}^{\infty} [x^{k+1}/(k + 1)] + C\right\}. \tag{14}$$

The constant term on the left side of (14) is 0, and hence the constant of integration

C is also 0. Next we multiply both sides of (14) by $1/(1-x)^2$:

$$y = [2/(1-x)^2]\sum_{k=1}^{\infty}[x^{k+1}/(k+1)] = 2\left(\sum_{k=1}^{\infty}kx^{k-1}\right)\left(\sum_{k=1}^{\infty}[x^{k+1}/(k+1)]\right)$$
$$= 2\sum_{n=2}^{\infty}x^n\sum_{k=2}^{n}[(n-k+1)/k]. \tag{15}$$

It is easily verified that this is the only power series about the origin whose constant term and coefficient of x are zero and which is a solution of the differential equation (11). Therefore it must be also $d(x)$. Hence we have the following explicit formula for $d_n/n!$:

$$d_n/n! = 2\sum_{k=2}^{n}[(n-k+1)/k]. \tag{16}$$

On dividing this equation by n, its right side can be manipulated routinely into the form of eq. (1). Note that a consequence of (14) is that $d(x)$ can also be expressed as

$$d(x) = 2(-x - \log(1-x))/(1-x)^2, \tag{17}$$

a fact that will be useful later.

Now we shall compare the mean distance over all sequences with the average distances in the highly balanced tree of order $n = 2^m - 1$ which has all of its points within a distance of $m-1$ to the root. Figure 3 displays such a tree with $m = 4$.

The sum s_n of distances from the root to all other points is

$$s_n = \sum_{k=1}^{m-1}k\cdot 2^k. \tag{18}$$

On subtracting $2s_n$ from s_n, we find

$$s_n = 2 + (m-2)2^m, \tag{19}$$

and therefore the average distance is

$$s_n/n = \log_2 n - 2 + R_n', \tag{20}$$

where $\lim_{n\to\infty} R_n' = 0$.

For intermediate values of n between $2^m - 1$ and 2^{m+1}, $\log_2 n - 2$ also underestimates only slightly (by less than .1) the average distance in a best possible tree. On comparing this result with formula (3) of Corollary 1, we see that the mean for all sequences is only about 40 percent higher than that of a perfect tree.

Several methods have been suggested for rearranging the stored data for the purpose of reducing the number of comparisons required to locate one item of information. On the basis of the information in this section we can formulate a criterion for deciding when such a rearrangement is uneconomical. Suppose $f(n)$ is the number of comparisons required to rearrange n items so that the new arrangement has the most favorable average search time of $\log_2 n - 2$. Then, on the average, the number of comparisons required to find t items is $f(n) + t(\log_2 n - 2)$. But from Corollary 1 we know that without rearranging the data, this same average is $t(1.4 \log_2 n - 2.8)$. Evidently the data should not be

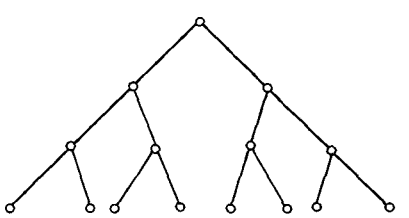

FIG. 3. The balanced tree of order $2^4 - 1$

rearranged if the number t of items to be found satisfies the inequality:

$$t < f(n)/(.4 \log_2 n - .8). \tag{21}$$

3. *Variance of the Mean Number of Comparisons*

We continue to consider the random variable which assigns to each sequence the mean distance of the points from the root in its associated binary tree. To determine the variance of this distribution we require σ_n, the sum over all sequences $a_1, \cdots, a_n$ of the square of the sum of the distances to the root of the points in the tree associated with $a_1, \cdots, a_n$.

THEOREM 2. *The variance of the means of all $n!$ sequences of n different integers is*

$$\sigma_n/n!n^2 - (d_n/n!n)^2 = 7 - 4(1 + 1/n)^2 \sum_{k=1}^{n} (1/k^2) + 13/n - 2(n+1)/n^2 \sum_{k=1}^{n} (1/k). \tag{22}$$

This expression for the variance can be simplified by using the fact that

$$\sum_{k=1}^{n} (1/k^2) = \pi^2/6 + O(1/n). \tag{23}$$

COROLLARY 2. *The variance of the mean number of comparisons is*

$$\sigma_n/n!n^2 - (d_n/n!n)^2 = .4202637 \cdots + R_n'', \tag{24}$$

where $.4202637 \cdots = 7 - 2\pi^2/3$, $|R_n''| < \frac{1}{2}$ *for all* n, *and* $R_n'' \sim (-2 \log n)/n$.

PROOF OF THE THEOREM. It is elementary that the variance of a distribution is the mean of the squares minus the square of the mean. The mean of the tree means is the number $d_n/n!n$, determined in Theorem 1. The mean of the squares of these tree means is $\sigma_n/n!n^2$.

To find the right-hand side of (22) we start with a recurrence relation for σ_n. As in the proof of Theorem 1, the contributions to σ_n due to sequences $a_1, \cdots, a_n$ with $a_1 = p$ are computed first, and the result is then summed over $p = 1, \cdots, n$. For any sequence the total distance in the corresponding tree is given by (5), so the square of this distance is $A^2 + B^2 + 2AB + 2(n-1)A + 2(n-1)B + (n-1)^2$. This leads to the relation

$$\sigma_n = \sum_{p=1}^{n} \binom{n-1}{p-1} [(n-p)!\sigma_{p-1} + (p-1)!\sigma_{n-p} + 2d_{p-1}\, d_{n-p} + 2(n-1)(n-p)!\, d_{p-1} + 2(n-1)(p-1)!\, d_{n-p} + (n-1)^2(p-1)!(n-p)!], \tag{25}$$

since the sum of A^2 over all $(p-1)!$ possible orderings of the first subsequence is σ_{p-1}, the sum of A is just d_{p-1}, and similarly for B^2 and B. The factor $\binom{n-1}{p-1}$ is again the number of ways of interlacing the subsequence of numbers less than p with the others.

Dividing both sides of (25) by $(n-1)!$ and simplifying gives

$$n\sigma_n/n! = \sum_{p=1}^{n} [\sigma_{p-1}/(p-1)! + \sigma_{n-p}/(n-p)! + 2d_{p-1}\, d_{n-p}/(p-1)!(n-p)! + 2(n-1)(d_{p-1}/(p-1)! + d_{n-p}/(n-p)!) + (n-1)^2]. \tag{26}$$

This again suggests the use of exponential generating functions, so let

$$\sigma(x) = \sum_{n=2}^{\infty} \sigma_n x^n/n!. \tag{27}$$

Now by multiplying both sides of (26) by x^{n-1} and summing over $n = 2, 3, \cdots$ we obtain

$$\sigma'(x) = 2\sigma(x)/(1-x) + 2d(x)^2 + 4x[d(x)/(1-x)]' + (2x + 4x^2)/(1-x)^4. \tag{28}$$

A more convenient form of (28) is found by substituting the closed form (17) for $d(x)$, giving

$$\sigma'(x) - 2\sigma(x)/(1-x) = (1-x)^{-4}[2x - 4x^2 - 8x\log(1-x) + 8\log^2(1-x)]. \quad (29)$$

As before, this equation can be solved by multiplying by the integrating factor $(1-x)^2$ and integrating the right side by parts. The result is

$$\sigma(x) = (1-x)^{-3}[2x + 4x^2 + 2(1+3x)\log(1-x) + 4(1+x)\log^2(1-x)]. \quad (30)$$

To determine $\sigma_n/n!$ from this equation, we need the next two equations which are routinely derived:

$$(1-x)^{-3}(2x+4x^2) = \sum_{n=1}^{\infty}(3n^2-n)x^n, \quad (31)$$

$$2(1-x)^{-3}\log(1-x) = -\sum_{n=1}^{\infty}x^n\{(n+1)(n+2)\left(\sum_{k=1}^{n}(1/k)\right) - n(3n+5)/2\}. \quad (32)$$

On squaring $\log(1-x)$ we find

$$\log^2(1-x) = \sum_{k=2}^{\infty}x^k\sum_{i=1}^{k-1}(1/i(k-i)) = \sum_{k=2}^{\infty}(x^k/k)\sum_{i=1}^{k-1}(1/i + 1/(k-i)) = 2\sum_{k=2}^{\infty}(x^k/k)\sum_{i=1}^{k-1}(1/i). \quad (33)$$

On multiplying this series by $(1-x)^{-3}$ as given in (10), we have

$$(1-x)^{-3}\log^2(1-x) = \sum_{n=2}^{\infty}x^n\left\{\sum_{m=0}^{n-2}(m+1)(m+2)/(n-m)\sum_{i=1}^{n-m-1}(1/i)\right\}. \quad (34)$$

On substitution in the right side of (34) with the identity

$$(m+1)(m+2)/(n-m) = (n+1)(n+2)/(n-m) - (n+3+m), \quad (35)$$

we obtain

$$(1-x)^{-3}\log^2(1-x) = \sum_{n=2}^{\infty}x^n\left\{(n+1)(n+2)\sum_{k=2}^{n}(1/k)\sum_{i=1}^{k-1}(1/i) - (n+3)\sum_{m=0}^{n-2}\sum_{m=0}^{n-m-1}(1/i) - \sum_{m=0}^{n-2}m\sum_{i=1}^{n-m-1}(1/i)\right\}. \quad (36)$$

The next two identities are easily derived by changing the order of summation:

$$\sum_{m=0}^{n-2}\sum_{i=1}^{n-m-1}(1/i) = n\left(\sum_{k=1}^{n}(1/k)\right) - n \quad (37)$$

$$\sum_{m=0}^{n-2}m\sum_{i=1}^{n-m-1}(1/i) = \binom{n}{2}\left(\sum_{k=1}^{n}(1/k)\right) + 3n(1-n)/4. \quad (38)$$

On substituting these values in eq. (36) above and simplifying, the result is

$$(1-x)^{-3}\log^2(1-x) = \sum_{n=2}^{\infty}x^n\left\{(n+1)(n+2)\sum_{k=2}^{n}(1/k)\sum_{i=1}^{k-1}(1/i) + (7n^2+9n)/5 - [(3n^2+5n)/2]\sum_{k=1}^{n}(1/k)\right\}. \quad (39)$$

The three equations (31), (32), and (39) can now be used with (30) to determine the following formula for $\sigma_n/n!$:

$$\sigma_n/n! = 23n^2 + 13n - 2(n+1)(8n+1)\sum_{k=1}^{n}(1/k) + 8(n+1)^2\sum_{k=2}^{n}(1/k)\sum_{i=1}^{k-1}(1/i). \quad (40)$$

Finally, we divide the right side of (40) by n^2, subtract $(d_n/n!n)^2$, and substitute using the equation

$$\sum_{k=1}^{n}(1/k)\sum_{i=1}^{n}(1/i) = 2\sum_{k=2}^{n}(1/k)\sum_{i=1}^{k-1}(1/i) + \sum_{k=1}^{n}(1/k^2), \tag{41}$$

to obtain formula (22) of Theorem 2.

4. *Variance of the Distance of a Random Point to the Root*

Let v_n be the sum of the squares of the distances to the root of each of the n points in each of the rooted binary trees determined by the $n!$ sequences $a_1, \cdots, a_n$.

THEOREM 3. *The variance of the distance from the root of the $n!n$ points in the binary trees determined by all $n!$ sequences $a_1, \cdots, a_n$ of n different integers is*

$$v_n/n!n - (d_n/n!n)^2 = 2(1 + 5/n)\sum_{k=1}^{n}(1/k) + 4 - 4(1 + 1/n)\sum_{k=1}^{n}(1/k^2) - 4(1/n + 1/n^2)\left(\sum_{k=1}^{n}(1/k)\right)^2. \tag{42}$$

PROOF. Since the proof follows the same lines as the proof of Theorem 2, we shall confine it to a sketch of the main points. The analogue of (25) is the recurrence relation

$$v_n = \sum_{p=1}^{n}\binom{n-1}{p-1}[(n-p)!\,v_{p-1} + (p-1)!\,v_{n-p} + 2(n-p)!\,d_{p-1} + 2(p-1)!\,d_{n-p} + (n-1)(p-1)!(n-p)!]. \tag{43}$$

This can be rewritten as

$$nv_n/n! = \sum_{p=1}^{n}[v_{p-1}/(p-1)! + v_{n-p}/(n-p)! + 2d_{p-1}/(p-1)! + 2d_{n-p}/(n-p)! + (n-1)]. \tag{44}$$

In terms of the exponential generating function $d(x)$ and

$$v(x) = \sum_{n=2}^{\infty} v_n x^n/n!, \tag{45}$$

eq. (44) takes the form

$$v'(x) = 2v(x)/(1-x) + 4d(x)/(1-x) + 2x/(1-x)^3. \tag{46}$$

This differential equation is solved in the same fashion as (11) with the following result:

$$v(x) = \{6x + 6\log(1-x) + 4\log^2(1-x)\}/(1-x)^2, \tag{47}$$

or what is the same,

$$v(x) = -3d(x) + 4\log^2(1-x)/(1-x)^2. \tag{48}$$

To determine $v_n/n!$ from (48), we multiply the right side of (33) by $(1-x)^{-2}$ as expressed in (9) and we have

$$(1-x)^{-2}\log^2(1-x) = 2\sum_{n=2}^{\infty}x^n\sum_{m=0}^{n-2}[(m+1)/(n-m)]\sum_{i=1}^{n-m-1}(1/i). \tag{49}$$

Since $(m+1)/(n-m) = (n+1)/(n-m) - 1$, this equation can be written

$$\begin{aligned}(1-x)^{-2}\log^2(1-x) &= 2\sum_{n=2}^{\infty}x^n\left[(n+1)\left\{\sum_{m=0}^{n-2}1/(n-m)\sum_{i=1}^{n-m-1}(1/i)\right\} - \sum_{m=0}^{n-2}\sum_{i=1}^{n-m-1}(1/i)\right] \\ &= 2\sum_{n=2}^{\infty}x^n\left[(n+1)\left\{\sum_{k=1}^{n}(1/k)\sum_{i=1}^{k-1}(1/i)\right\} - \sum_{k=1}^{n-1}[(n-k)/k]\right]\end{aligned}$$

$$= 2 \sum_{n=2}^{\infty} x^n \left[(n+1) \left\{ \sum_{k=2}^{n} (1/k) \sum_{i=1}^{k-1} (1/i) \right\} + (n-1) - n \sum_{k=1}^{n-1} (1/k) \right]. \quad (50)$$

Therefore from this equation and (48) we have

$$v_n/n!n = -3d_n/n!n + 8(1 + 1/n) \sum_{k=2}^{n} (1/k) \sum_{i=1}^{k-1} \tfrac{1}{8} + 8 - 8 \sum_{k=1}^{n} (1/k). \quad (51)$$

The variance given in (42) can now be calculated using formula (1) of Theorem 1 for d_n and the identity (41). The terms in (42) which go to zero can be collected by referring to (2) and (23), with the result

$$v_n/n!n - (d_n/n!n)^2 = 1.386 \log_2 n - 1.425 + O(\log^2 n/n). \quad (52)$$

Here $1.386 \cdots = 2 \log 2$ and $1.425 \cdots = 4(\pi^2/6 - 1) - 2\gamma$. Note the similarity of (52) to (3).

5. *Summary*

The results of Sections 2, 3, and 4 provide a first-order description of the distribution of comparison times in binary search trees. Each tree determines its own mean and variance of comparison times; call these the tree-mean and the tree-variance. Then the mean of the tree-means over all binary search trees of order n is just the average $d_n/n!n$ determined in Theorem 1 to be asymptotic to $2 \log n$. The variance of these tree-means was found in Theorem 2 to approach a small constant as a limit. Intuitively, this implies that the chance of picking at random a search tree of order n with a significantly different tree-mean from the average becomes small as n increases. More precisely, a tree of order n with mean μ would be said to differ from the average by a factor of ϵ just if $|\mu - d_n/n!n| \geq \epsilon d_n/n!n$. It follows from Theorems 1 and 2 that for any $\epsilon > 0$ the probability of picking at random a binary search tree of order n with tree-mean differing from the average by a factor of ϵ approaches zero more quickly than $1/\log^2 n$.

Information on the rate of growth of the mean of the tree-variances, over all trees of order n, is implicit in Theorems 2 and 3. For the overall variance of the distance of a random point to the root of a random tree of order n (computed in Theorem 3) equals the sum of the variance of the tree-means (computed in Theorem 2) and the mean of the tree-variances. Thus the mean of the tree-variances is asymptotic to $2 \log n$, and thus accounts for most of the variance computed in Theorem 3.

There is little doubt that higher moments of the comparison time distributions can be calculated by the methods already presented. In the light of recent developments, however, such computations appear to be of minor significance.

Briefly, the new results consider a different premise about the nature of the binary comparisons by which the search trees are constructed. Instead of a linear order we start with a random binary relation. This leads to a substantial reduction in the average number of comparisons required to retrieve an item from a search tree. These results will be the subject of a separate paper.

REFERENCES

1. HIBBARD, T. N. Some combinatorial properties of certain trees with applications to searching and sorting. *J. ACM 9* (1962), 13–28.
2. MARTIN, W. A., AND NESS, D. N. Optimizing binary tree growth with a sorting algorithm. *Comm. ACM 15*, 2 (Feb. 1972), 88–93.
3. WINDLEY, P. F. Trees, forests and rearranging. *Computer J. 3* (1960), 84–88.

RECEIVED MAY 1973; REVISED SEPTEMBER 1973