SOME OBSERVATIONS ON RANDOM 2-3 TREES

Mark R. BROWN *

Department of Computer Science, Yale University, New Haven, CT 06520, U.S.A.

Received 20 April 1979

Analysis of algorithms, 2-3 trees

1. Introduction

Balanced tree structures such as 2-3 trees [1,3] are a method of organizing information that allows both fast accessing and fast updating. For example, 2-3 trees may be used to represent arbitrary linear lists of length n such that items can be inserted into and deleted from a list in O(log n) time.

A precise analysis of the average-case behavior of 2-3 trees seems to be quite difficult. Yao [5] gives a method that, in principle, can analyze any fixed number of levels at the bottom of a 2-3 tree, as the tree size tends to infinity. He actually performs the analysis for the bottom two levels, noting that the calculations become substantially more complicated in the three level case and seemingly impossible for four or more levels.

In this note we report the results of a three level analysis of random 2-3 trees, using Yao's method. We also give the results of simulation experiments that shed some light on the behavior of higher levels in random 2-3 trees. We assume that the reader is familiar with 2-3 trees and with Yao's analysis in what follows.

2. Three level analysis

Yao's one level analysis involved two subtree types, and the two level analysis involved seven types. There are 978 distinct 2-3 trees of height three, where trees that are mirror-images are considered to be identical. Thus a three level analysis can be carried out using a 978 X 978 matrix D [5, eq. 5]. (It is likely that an equivalent analysis can be carried out on a somewhat smaller matrix, by grouping trees into less refined classes; Yao reduced the two level D matrix from 12 X 12 to 7 X 7 in this way.) This D matrix, whose entries are determined by the results of all possible insertions into each of the 978 trees, was generated using a computer program. The matrix is sparse, containing only 9621 nonzero elements. This made it possible to verify numerically that (D – I) satisfies the conditions of Yao's Lemma 2.10, and to determine the 978 component vector u that gives the asymptotic solution. As a consistency check, the three level solution was used to derive the two level solution given by Yao.

A complete listing of the solution vector here would be unenlightening, so instead we shall present certain statistics gathered from the solution. First, some notation: a *2-node* is an internal node of a 2-3 tree with 2 children, and a *3-node* is a similar node with 3 children. *Level* 1 of a 2-3 tree is the bottommost level of internal nodes (the internal nodes farthest from the root); the parents of these nodes lie in *level* 2, and so on. Let $A_2(l, N)$ denote the expected number of 2-nodes on level l of a random 2-3 tree with N external nodes (N – 1 keys), and define $A_3(l, N)$ similarly for 3-nodes. Finally, let $\rho(l, N) = A_2(l, N)/A_3(l, N)$, that is, the ratio of the expected numbers of 2- and 3-nodes on level l.

Table 1 gives the asymptotic values (as $N \to \infty$) of

* This work was partially supported by the National Science Foundation under Grant MCS-7813764.

Table 1
Composition of the bottom three levels of a random 2-3 tree

l	$(1/N) \cdot A_2(l, N)$	$(1/N) \cdot A_3(l, N)$	$\rho(l, N)$
3	$0.0502777 + O(N^{-4.37})$	$0.0271748 + O(N^{-4.37})$	$1.8501612 + O(N^{-4.37})$
2	$0.1176681 + O(N^{-6.55})$	$0.0644117 + O(N^{-6.55})$	$1.8268110 + O(N^{-6.55})$
1	$0.2857143 + O(N^{-7})$	$0.1428571 + O(N^{-7})$	$2 \qquad + O(N^{-7})$

$A_2(l, N)$, $A_3(l, N)$, and $\rho(l, N)$ for $l \leqslant 3$. It may be surprising that ρ is not monotonic in l.

3. Empirical tests

We performed simulation experiments to gain some information about the higher levels of 2-3 trees. The experiments consisted of inserting a sequence of independent random keys into an initially empty 2-3 tree. The keys were integers generated by an additive random number generator of the type analyzed by Reiser [4].

Table 2 gives statistics gathered from 2000 random 2-3 trees with 30 000 keys each. The confidence intervals are estimated from the observed standard deviations in the 2000 observations. One potential difficulty in interpreting the results is that we want asymptotic values but our trees are finite; we expect that finiteness may have a large influence on the composition of levels high in the trees. To gauge this influence, we performed experiments for different tree sizes and compared the results. The statistics for levels shown in the table did not change substantially when the tree size was increased or decreased by a factor of two.

The figures in Table 2 agree well with the analytical results of Table 1 for the bottom three levels. In higher levels, for which no analysis has been performed, it appears that the proportion of 2-nodes increases steadily as we go higher in the tree.

4. Conclusions

Many interesting questions about 2-3 trees remain unanswered. The empirical tests suggest that the composition of higher levels of a random 2-3 tree is quite different from that of the three lower levels that have been analyzed. It is even possible that $\rho(l)$ is unbounded as $l \to \infty$, but there is no analytical evidence either way on this question.

It would be interesting to find even an approximate analytical model for the higher levels of 2-3 trees. One natural approach is to view the bottom l levels of a 2-3 tree as a 'filter' that determines the pattern of insertions into the $l + 1^{st}$ level. If this filter has the property that the insertions coming out of it are independent, then a Yao-style analysis may be able to derive the characteristics of an $l + 1$ level filter from those of an l level filter. But it appears that the true state of affairs is more complicated –

Table 2
Composition of the bottom eight levels of a random 2-3 tree with N = 30 001, estimated by empirical tests

l	$(1/N) \cdot A_2(l, N)$	$(1/N) \cdot A_3(l, N)$	$\rho(l, N)$
8	0.000786 ± 0.000005	0.000354 ± 0.000004	2.220 ± 0.039
7	0.001798 ± 0.000008	0.000834 ± 0.000006	2.156 ± 0.025
6	0.004114 ± 0.000013	0.001983 ± 0.000008	2.075 ± 0.015
5	0.00946 ± 0.00002	0.00472 ± 0.00001	2.0042 ± 0.0085
4	0.02177 ± 0.00003	0.01130 ± 0.00002	1.9265 ± 0.0061
3	0.05025 ± 0.00005	0.02719 ± 0.00003	1.8481 ± 0.0039
2	0.11764 ± 0.00007	0.06443 ± 0.00005	1.8259 ± 0.0025
1	0.28573 ± 0.00010	0.14285 ± 0.00007	2.0000 ± 0.0017

insertions into higher levels of the tree are highly nonindependent. An analysis that accounts for this lack of independence seems to be necessary.

Analysis of the number of 2- and 3-nodes on each level does not furnish answers to other interesting questions about 2-3 trees. Since 2-3 trees are often represented as binary trees [3], we would like to analyze the expected path length of the binary representation for a 2-3 tree. Guibas and Sedgewick [2] have shown that such path length questions can be answered for algorithms that never perform rebalancings above a fixed level in the tree, but no known algorithm with a O(log n) worst case has this property. If we could show that rebalancings above a fixed [illegible] the tree do not (on the average) increase the tree's path length, then the technique of Guibas and Sedgewick would give greatly improved bounds on the path length of random 2-3 trees.

Acknowledgment

I wish to thank Michael Merritt for writing the program to generate the D matrix, and Dan Boley and Stan Eisenstat for helping with numerical problems.

References

[1] A.V. Aho, J.E. Hopcroft and J.D. Ullman, The Design and Analysis of Computer Algorithms (Addison-Wesley, Reading, MA, 1974).

[2] L.J. Guibas and R. Sedgewick, A dichromatic framework for balanced trees, Proc. 19th Annual Symposium on Foundations of Computer Science, Ann Arbor, MI (1978) 8–21.

[3] D.E. Knuth, The Art of Computer Programming, Vol. 3: Sorting and Searching (Addison-Wesley, Reading, MA, 1973) Section 6.2.3.

[4] J.F. Reiser, Analysis of additive random number generators, Stanford University Computer Science Department Report STAN-CS-77-601 (1977).

[5] A.C. Yao, On random 2-3 trees, Acta Informat. 9 (1978) 159–170.