RANDOM INSERTIONS IN TREES AND RELATED TOPICS

DISSERTATION SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (TECHNOLOGY) IN COMPUTER SCIENCE OF THE UNIVERSITY OF CALCUTTA

ASIM KUMAR PAL
INDIAN INSTITUTE OF MANAGEMENT CALCUTTA
P. O. BOX 16757, CALCUTTA-700027, INDIA

1984

1.2 Summary of Thesis

This thesis concerns itself mainly with random insertions in balanced and unbalanced tree structures. One of the objectives is to compare different tree structures by deriving bounds on various performance measures. The other interest is primarily mathematical; most of the analysis involve the formulation and solution of linear recurrence relations with variable coefficients, and the solutions are not always routine, sometimes even requiring the use of asymptotic techniques.

The thesis is based on the four papers [5], [6], [39] and [40]. There are five chapters. The first chapter is introductory. The contents of the others are summarized below.

The subject matter of Chapter Two is the generalized Polya-Eggenberger urn model. In this model, an urn initially contains a given number of white and black balls. A ball is selected at random from the urn, and the number of white and black balls added to (or taken away from) the urn depends on the colour of the ball selected. Let w_n be the random variable giving the number of white balls in the urn after n draws. A sufficient condition is derived for the asymptotic normality, as $n \longrightarrow \infty$, of the standardized random variable corresponding to w_n. This result can be used to obtain the distribution, under random insertions, of binary and ternary nodes at the lowest level in a 2-3 tree, thereby extending the first-order analysis of Yao [45], who looked only at expected values.

Chapter 3 tries to develop, in a very general framework, a Markov chain model of the random insertion process in tree structures. The Markov chain is nonhomogeneous, and the properties of the transition matrix need to be carefully scrutinized in order to find the limiting probability vector.

This chapter can be viewed as an interesting application of matrix algebra methods.

A new data structure, called a 3-tree, gets introduced in Chapter 4. A 3-tree is a tree in which every node has either 2 or 3 sons, where some or all of the successors may be null. 2-3 trees are special cases of 3-trees. A simple search-and-insertion algorithm for 3-trees is first formulated. The properties of the 3-trees that get built by this algorithm are then analyzed assuming that insertions are random. It is shown that the number of key comparisons for a successful or an unsuccessful search is the same as in the binary case. Various properties of these randomly constructed 3-trees are then compared with corresponding properties of binary trees and of different types of 2-3 trees. To give an idea about the expected "shape" of a random 3-tree, the notion of the weight balance factors of a node is introduced, and the distribution of balances of the keys associated with the nodes of a large randomly constructed 3-tree is derived.

Chapter 5 defines a new performance measure for binary trees called the mean weight balance factor (MWBF). For any binary tree T, $0 < MWBF(T) \leq 1$. Very unbalanced trees have MWBF close to 0, while complete binary trees have MWBF close to 1. The expected MWBF of a binary tree under random insertion is derived. It is shown that an AVL tree has an MWBF of

atleast 0.73. Bounds are also obtained on the expected MWBF of an AVL tree under random insertions.

Further work is possible in many of the above areas. Problems that remain open are described in the respective chapters. One important issue remains largely unresolved. No totally satisfactory analysis of random deletions in tree structures exists as yet; preliminary attempts have been reported by Knuth [33], Mehlhorn [35] and Jonassen and Knuth [28], while Eppinger [15] has made some empirical observations. This thesis does not have anything to say on random deletions, but it is earnestly hoped that its contents will encourage others to take a fresh look at the deletion problem.

Since each chapter of the thesis is essentially self-contained and does not use results from other chapters, it has been found convenient to have separate lists of notation for the different chapters.