Acta Informatica (2006) 43:243–264
DOI 10.1007/s00236-006-0019-7

ORIGINAL ARTICLE

Distances in random digital search trees

Rafik Aguech · Nabil Lasmar · Hosam Mahmoud

Received: 29 July 2005 / Revised: 12 April 2006 /
Published online: 22 September 2006

Abstract Distances between nodes in random trees is a popular topic, and several classes of trees have recently been investigated. We look into this matter in digital search trees. By analytic techniques, such as the Mellin Transform and poissonization, we describe a program to determine the moments of these distances. The program is illustrated on the mean and variance. One encounters delayed Mellin transform equations, which we solve by inspection. In addition to various asymptotics, we give an exact expression for the mean and for the variance in the unbiased case. Interestingly, the unbiased case gives a bounded variance, whereas the biased case gives a variance growing with the number of keys. It is therefore possible in the biased case to show that an appropriately normalized version of the distance converges to a limit. The complexity of moment calculation increases substantially with each higher moment; it is prudent to seek a shortcut to the limit via a method that avoids the computation of all moments. Toward this end, we utilize the contraction method to show that in biased digital search trees the distribution of a suitably normalized version of the distances approaches a limit that is the fixed-point solution of a distributional equation (distances being measured in the Wasserstein metric space). An explicit solution to the fixed-point equation is readily demonstrated to be Gaussian.

Keywords Random trees · Recurrence · Mellin transform · Poissonization · Fixed point · Contraction method

R. Aguech
Département de mathématiques, Faculté des Sciences de Monastir, 5019 Monastir, Tunisia
e-mail: rafikaguech@ipeit.rnu.tn

N. Lasmar
Département de mathématiques, Institut préparatoire aux études d'ingénieurs de Tunis, IPEIT, Rue lelnahrou-Montfleury, Tunis, Tunisia
e-mail: nabillasmar@yahoo.fr

H. Mahmoud (✉)
Department of Statistics, The George Washington University, Washington, DC 20052, USA
e-mail: hosam@gwu.edu

AMS Subject Classifications Primary: 05C05 · 60C05; secondary: 60F05 · 68P05 · 68P10 · 68P20

1 Introduction

The behavior of distances between nodes in random trees has lately become a topic of interest, as can be seen in half a dozen or so of recent papers. Neininger [21] looked at these distances in recursive trees, Mahmoud and Neininger [18] studied these distances in binary search trees. The method used in these two papers was contraction. Devroye and Neininger [5] revisited the subject of binary search trees and refined the results using more elementary arguments. Their paper also takes up several other types of distances not considered in [18]. Employing generating functions and ensuing functional equations, Panholzer and Prodinger [23] generalized the result to the size of spanning trees for a randomly chosen set of nodes. Christophi and Mahmoud [3] considered the unbiased "tries" and demonstrated that the distribution of distances is oscillatory, with no possible nontrivial limit. Aguech, Lasmar and Mahmoud [1] considered the contrast that one encounters in biased tries, showing the existence of a Gaussian limit for a (normalized) version of the distance. A disparate variety of methods has been used in these studies.

Distances in yet another natural class of digital trees remain uninvestigated. It is the class of digital search trees, a class similar to tries, but the keys are kept in the nodes, instead of using them only as indexes for branching as in tries. The construction algorithm of the digital search tree seems more natural than that of the trie, and has an advantage of putting an upper bound on the size of the tree, whereas tries can degenerate into near linear structures with very long paths (with low probability of course).

We wish to study distances in a digital search tree growing on n keys. Some tough recurrences appear in the study and seem to be very challenging. The idea of poissonization is well-suited for such a recurrence. The method has become a popular tool. It involves a poissonization-Mellin-inverse Mellin-depoissonization program. The method is beginning to appear as a chapter on standard techniques in information science; see [31] for example, and numerous references within. Broadly speaking, the program works as follows. If a Poisson number of keys is assumed instead, the functional equations involved can asymptotically be solved by the Mellin transform and its inverse. The solution is a good approximation (with ignorable errors) for the fixed-population problem, when the Poisson parameter is taken to be n, as $n \to \infty$.

The complexity of such a Mellin approach increases considerably for higher moments, which strongly invites a consideration of some shortcut. The contraction method provides such a direct bridge to the limit.

The standard data model for digital search trees is the Bernoulli probability distribution, which should ideally be unbiased. In practice this unbias is not guaranteed in view of the sensitivity of data generators. So, our study is not limited to the unbiased Bernoulli case, and puts in good perspective the contrast between biased and unbiased data models. We assume that keys of infinite precision are obtained from a memoryless source that emits independent bits, with $\mathbf{P}(\text{Bit} = 1) = p$, and $\mathbf{P}(\text{Bit} = 0) = q = 1 - p$. We say the Bernoulli model (or the resulting digital search tree) is unbiased if $p = q = \frac{1}{2}$, otherwise it is biased. It is desirable, but not guaranteed,