Some results about quadtrees

Louise Laforest
UQAM, Montréal

June 7, 1993

[summary by Bruno Salvy]

Introduction

The quadtree data structure is a natural generalization of binary search trees in higher dimension used to store multidimensional data. The average-case complexity of algorithms operating on quadtrees is directly related to the expectation of some parameters such as path length, number of leaves, number of nodes with k children, and so on.

Parameters studied here are *additive parameters* that can be computed recursively by adding a toll number p_n depending only on the size n of the tree to the values of the parameter on the subtrees. For instance the number of leaves is obtained with $p_n = \delta_{n,1}$. The expectation of such a parameter e_n obeys the following classical recurrence [2, 4]

$$e_n = p_n + 2^d \sum_{k=0}^{n-1} \pi_{n,k} e_k, \tag{1}$$

where d is the dimension ($d = 2$ for standard quadtrees) and $\pi_{n,k}$ is the probability that a quadtree of size n has its first subtree of size k. This probability is given by the following explicit formula [1, 2, 4]

$$\begin{aligned} \pi_{n,k} &= \frac{1}{n} \sum_{k < i_1 \le i_2 \le \cdots \le i_d \le n} \frac{1}{i_1 \cdots i_n}, \\ &= \binom{n-1}{k} \sum_{j=0}^{n-1-k} \binom{n-1-k}{j} \frac{(-1)^j}{(k+j+1)^d}. \end{aligned}$$

In dimension 2, this reduces to

$$\pi_{n,k} = \frac{1}{n}[H_n - H_k],$$

H_n denoting the n-th harmonic number.

Taking generating functions, formula (1) can be seen to be equivalent to a simple integral equation,

$$e(z) = p(z) + 2^d \mathbf{J}^{d-1} \mathbf{I} e(z), \tag{2}$$

where $\mathbf{J}$ and $\mathbf{I}$ are the following integral operators

$$\mathbf{I}f(z) = \int_0^z f(t) \frac{dt}{1-t}, \quad \mathbf{J}f(z) = \int_0^z f(t) \frac{dt}{t(1-t)}.$$

Equation (2) is easily translated into a linear differential equation of order d. Combined with singularity analysis, this differential equation is the basis of most of the analysis in [2].

1. A simplifying remark

At this stage, G. Labelle and L. Laforest [3] introduce the following change of variable and unknown function:

$$e(z) \mapsto e^\star(z) = (1-Z)e(Z), \quad Z = 1 - \frac{1}{1-z}.$$

This transformation is involutive, i.e., $e^{\star\star} = e$. Under this transformation, Equation (2) is translated into

$$(1-Z)e^\star(Z) = (1-Z)p^\star(Z) + 2^d \mathbf{J}^{d-1}\mathbf{I}(1-Z)e^\star(Z).$$

The integral operators $\mathbf{I}$ and $\mathbf{J}$ have a nice action on the variable $Z(z)$:

$$\mathbf{I}(1-Z(z))f(Z(z)) = -\int_0^{Z(z)} f(u)\,du, \quad \mathbf{J}f(Z(z)) = \int_0^{Z(z)} f(u)\frac{du}{u},$$

so that their action on the Taylor expansion of f *in the variable* Z is particularly simple: the first one negates the coefficients and shift them by 1 while the second one divides the nth Taylor coefficient by n. Taking coefficients of Z^n on both sides of the above equation thus yields *a linear recurrence of order 1*,

$$e_n^\star - e_{n-1}^\star = p_n^\star - p_{n-1}^\star - 2^d \frac{e_{n-1}^\star}{n^d}. \tag{3}$$

2. Characteristic constants

The real object of study is of course e_n for various p_n. Consider $\lambda_{\nu,d}$ defined as the limit of e_n/n when $p_n = \delta_{n,\nu}$ and n tends to infinity. The parameter e_n then records the number of trees of size ν in a tree of size n. These constants play a role in the analysis of storage allocation strategies. They can also be used to derive other asymptotic behaviours by linear combination.

Applying twice the involutive transformation mentioned above, plus manipulations on recurrence (3), the theorem is that

$$e(z) = p(z) + \int_0^1 K_d(z,t)p(tz)\,dt,$$

where K is a polynomial of degree $d-1$ in $\log t$ whose coefficients are power series in z and t. Unfortunately, these power series have known sums only for $d = 1, 2$. From these sums one deduces

$$\lambda_{\nu,1} = \frac{2}{(\nu+1)(\nu+2)}, \quad \lambda_{\nu,2} = 3 - 18\nu + 6\nu(3\nu+1)\psi'(\nu+1),$$

where ψ is the logarithmic derivative of Euler's Γ function. Taking $\nu = 1$ yields the expected asymptotic proportion of leaves in a 2-dimensional quadtree: $\lambda_{1,2} = 4\pi^2 - 39$.

The existence of a linear differential equation for the generating function makes it possible to compute values of e_n in linear time, so that the values of the characteristic constants can be numerically determined for large values of ν and small values of d.

Other constants can be obtained explicitly in dimension 2. For instance, the proportion of nodes having exactly one or two children are given respectively by

$$24\zeta(3) - 26\pi^2 + 228, \quad -132\zeta(3) + 24\pi^2\ln 2 + \frac{67}{2}\pi^2 - 336.$$

Bibliography

[1] Devroye (Luc) and Laforest (Louise). – An analysis of random d-dimensional quad trees. *SIAM Journal on Computing*, vol. 19, 1990, pp. 821–832.

[2] Flajolet (Philippe), Gonnet (Gaston), Puech (Claude), and Robson (J. M.). – Analytic variations on quadtrees. *Algorithmica*, no7, December 1993.

[3] Labelle (Gilbert) and Laforest (Louise). – Étude de constantes universelles pour les arborescences hyperquaternaires de recherche. In *Séries formelles et combinatoire algébrique.* – 1993. Proceedings of FPACS'5, Florence (Italy).

[4] Laforest (Louise). – *Étude des arbres hyperquaternaires.* – Technical Report n° 3, Montreal, LACIM, UQAM, November 1990. (Author's PhD Thesis at McGill University).