Distances from the root in skew trees.

by Edsger W. Dijkstra and C. S. Scholten

Renewed interest in algorithms for sorting in situ raised the question of the average distance from node to root in binary trees other than balanced ones. (Here binary trees are to be understood as rooted trees in which nodes have zero, one, or two sons.)

In particular we consider the infinite sequence of trees T_i $(i \geq 0)$, in which for some fixed p and q $(p > q \geq 0)$

1) T_i for $0 \leq i < p$ are arbitrarily chosen
2) T_n and T_{n+q} are the subtrees of T_{n+p}.

With H_i = the number of nodes in T_i, we have

$$H_{n+p} = H_{n+q} + H_n + 1 \quad ;$$

with G_i = the sum of the distances of the nodes of T_i from its root we have

$$G_{n+p} = G_{n+q} + G_n + H_{n+q} + H_n \quad .$$

We are interested in the asymptotic behaviour of G_i/H_i for large i, this ratio being the average distance from the root in T_i.

Without loss of generality we can confine ourselves to the case $\gcd(p,q) = 1$, since in the case

$\gcd(p,q) = g > 1$, the sequence T_i consists of an interleaving of g mutually independent sequences.

With A_i defined by $A_i = H_i + 1$ and B_i defined by $B_i = G_i - 2$, we derive

(0) $A_{n+p} = A_{n+q} + A_n$

(1) $B_{n+p} = B_{n+q} + B_n + A_{n+p}$.

Equation (1) is not homogeneous in the B's, but by solving it for the A's and substituting them in (0), we get a homogeneous recurrence relation for the B's, the characteristic polynomial of which is the product of (0)'s characteristic polynomial and the characteristic polynomial corresponding to the homogeneous part of (1). We conclude that the characteristic equation for the A_i is

(2) $$x^p - x^q - 1 = 0$$

and that that for the B_i is

(3) $$(x^p - x^q - 1)^2 = 0$$.

Under the constraints $\gcd(p,q) = 1$ and $p > q \geq 0$, (2) enjoys the property of having one positive root, r say, such that $r > 1$ and all other roots of (2) have a modulus smaller than r. (For a proof of this theorem, see later.)

From this and the theory of linear recurrence relations we conclude

1) that the leading term of A_i is of the form $k.r^i$ for some constant k
2) that the leading term of B_i is of the form $(\ell + L.i).r^i$ for some constants ℓ and L.

Substituting these leading terms in (1) we get

$$(\ell + L.(n+p)).r^{n+p} = (\ell + L.(n+q)).r^{n+q} + (\ell + L.n)r^n + k.r^{n+p}$$

Since r is a root of (2) this can be reduced to

$$(4) \qquad L/k = r^p/(p.r^p - q.r^q) \quad .$$

We define the skewness of a binary tree as the ratio (≥ 1) of the numbers of nodes in its two subtrees. For the trees T_i it follows from the leading term of A_i that the asymptotic skewness s is given by $s = r^q$, whence $q = {}^r\log s$. Remembering that r satisfies (2), we find $r^p = s+1$, whence $p = {}^r\log(s+1)$. Hence (4) can be rewritten as

$$(5) \qquad L/k = \frac{s+1}{(s+1).{}^r\log(s+1) - s.{}^r\log s} \quad .$$

Because $r>1$ — so that the asymptotic value of G_i/H_i equals that of B_i/A_i —, we conclude that, expressed in r and s, the average distance from the root in T_i is for large i

$$\frac{(s+1).i}{(s+1).{}^r\log(s+1) - s.{}^r\log s} \quad .$$

Consequently, the average distance from the root in a tree from the sequence T_i with N nodes is

$$\frac{(s+1).\log N}{(s+1).\log(s+1) - s.\log s} \quad ,$$

i.e. an expression in s and N only.

* * *

We are left with the obligation to prove that $f\,x = 0$ with $f\,x = x^p - x^q - 1$ has one positive root $r > 1$ dominating the others. Since $f\,0 = -1$ and $f(+\infty) = +\infty$, $f\,x = 0$ has an odd number of positive roots. Because

$$f'\,x = x^{q-1}.(p.x^{p-q} - q)$$

we conclude that $f'\,x = 0$ has at most 1 positive root; hence $f\,x = 0$ has 1 positive root r and, because $f\,1 = -1$, we conclude $r > 1$. In other words

(6) for $x \geq 0$ $\quad sgn(f\,x) = sgn(x - r)$.

In order to prove dominance of r, we consider a root $m.e^{i.\varphi}$ of (2), with $m > 0$. Consequently

$$m^p.e^{i.p.\varphi} = m^q.e^{i.q.\varphi} + 1 \quad ,$$

from which we derive – by taking absolute values –

$$m^p < m^q + 1 \quad \wedge \quad (p.\varphi) \underline{\bmod} 2\pi = (q.\varphi) \underline{\bmod} 2\pi = 0 \quad .$$

Since $\gcd(p,q)=1$, this can be rewritten as

$$m^p - m^q - 1 < 0 \quad \forall \quad \varphi = 0$$

or, in view of (6): $m < r \quad \forall \quad \varphi = 0$. q.e.d.

* * *

Finally we remark that thanks to

$$\frac{p}{q} = \frac{\log(s+1)}{\log s}$$

any value of s can be approximated arbitrarily closely by suitable choice of p and q , a fact that enhances the significance of the expression in s and N for the average distance from the root.

28 August 1981

drs. C.S. Scholten
Philips Research Laboratories
5600 MD EINDHOVEN
The Netherlands

prof. dr. Edsger W. Dijkstra
Burroughs Research Fellow
Plataanstraat 5
5671 AL NUENEN
The Netherlands