NOTE

EXTERNAL INTERNAL NODES IN DIGITAL SEARCH TREES VIA MELLIN TRANSFORMS

HELMUT PRODINGER

Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria

January 13, 1992

Knuth [4; ex.6.3.29] posed it as an open problem (ranked $M46$) to compute (asymptotically) l_N, the average number of external internal nodes in a random *digital search tree* built by N data.

This problem was settled by Flajolet and Sedgewick [2] by a very convenient technique, called *Rice's method.* This method has some advantages over the *Mellin transform technique* that Knuth used heavily in his famous book. (In passing it will be mentioned that the variance was treated by P. Kirschenhofer and the author [3].)

In this note we show that *Knuth could have given a solution himself*, because in the present approach we follow his ideas that he used in a similar (although easier) problem (cf. [4; p.497]).

For a complete description of the problem and digital search trees we must refer to [4] and [2] in order to keep this note short. For the use of the Mellin transform technique we cite [1].

The problem is to compute

$$l_N = N - \sum_{k=2}^{N} \binom{N}{k} (-1)^k R_{k-2}$$

with

$$R_N = Q_N \sum_{k=0}^{N} \frac{1}{Q_k}$$

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

and
$$Q_k = \left(1-\frac{1}{2}\right)\left(1-\frac{1}{4}\right)\dots\left(1-\frac{1}{2^k}\right) \quad \text{for } k\ge 1 \quad \text{and } Q_0=1.$$
It follows from [2] and [3] that
$$R_N = N+1-\alpha+R^*(N)$$
with
$$\alpha = \sum_{k\ge 1}\frac{1}{2^k-1} = 1.606695\dots$$
and
$$R^*(z) = \frac{1}{Q_\infty}\sum_{j\ge 1} a_j \frac{z+j}{2^{z+j}-1}.$$
The quantities in the last equation are
$$Q_\infty = \lim_{k\to\infty} Q_k = 2.88788\dots$$
and
$$a_j = (-1)^{j-1}2^{-\binom{j}{2}}/Q_{j-1}.$$
We mention for later use that
$$R^*(-1) = \frac{1}{Q_\infty}\left[\frac{1}{\log 2}+\sum_{j\ge 2} a_j \frac{j-1}{2^{j-1}-1}\right] = 2.2346464\dots$$
(a very fast converging series!).

Therefore
$$l_N \sim N(\alpha+1)-\Sigma$$
with
$$\Sigma = \frac{1}{Q_\infty}\sum_{j\ge 1} a_j \sum_{k=2}^{N}\binom{N}{k}(-1)^k\frac{k-2+j}{2^{k-2+j}-1},$$
and this sum Σ will be attacked in the style of Knuth: Expanding the geometric series we get
$$\begin{aligned}\Sigma &= \frac{1}{Q_\infty}\sum_{j\ge 1} a_j \sum_{m\ge 1}\sum_{k=2}^{N}\binom{N}{k}(-1)^k(k-2+j)2^{-m(k-2+j)}\\ &= \frac{1}{Q_\infty}\sum_{j\ge 1} a_j \sum_{m\ge 1}\Big[-N2^{-m(j-1)}\left[(1-2^{-m})^{N-1}-1\right]\\ &\qquad +(j-2)2^{-m(j-2)}\left[(1-2^{-m})^N-1+\frac{N}{2^m}\right]\Big].\end{aligned}$$

The last step was by the binomial theorem. Now let

$$\psi_1(x) = e^{-x} - 1 \qquad \text{and} \qquad \psi_2(x) = e^{-x} - 1 + x.$$

Then the sum Σ may be approximated by

$$\begin{aligned}\Sigma &\sim \frac{1}{Q_\infty} \sum_{j\ge 1} a_j \sum_{m\ge 1} \left[-\frac{N}{2^{m(j-1)}} \psi_1\left(\frac{N}{2^m}\right) + \frac{j-2}{2^{m(j-2)}} \psi_2\left(\frac{N}{2^m}\right) \right] \\ &=: \Sigma_1 + \Sigma_2.\end{aligned}$$

This is achieved by the *exponential approximation* $(1-a)^N \approx e^{-aN}$; see e.g. [4; p.131].

Let us start with Σ_1:

$$\Sigma_1 = -\frac{N}{Q_\infty} \cdot f_1(N)$$

with

$$f_1(x) = \sum_{j\ge 1} a_j \sum_{m\ge 1} 2^{-m(j-1)} \psi_1\left(\frac{x}{2^m}\right).$$

It's Mellin transform is for $-1 < \Re s < 0$

$$\begin{aligned} f_1^*(s) &= \sum_{j\ge 1} a_j \sum_{m\ge 1} 2^{-m(j-1)} 2^{ms} \Gamma(s) \\ &= \Gamma(s) \cdot \sum_{j\ge 1} a_j \frac{1}{2^{j-1-s} - 1}. \end{aligned}$$

The Mellin inversion formula

$$f(x) = \frac{1}{2\pi i} \int f^*(s) x^{-s}\, ds$$

(where the integration is along a vertical line inside the so called *fundamental strip*) allows to go back to the original function. To evaluate it asymptotically, we have to consider the negative residues of the integrand right of the fundamental strip (i.e. here at $s = 0$). We don't discuss the other residues which lead to smaller order terms resp. periodic fluctuations of small amplitude, but this can be done as easily as the main term.

The term $j = 1$ must be treated separately, as it contains a *second order pole*. The residue computation is standard and gives a contribution

$$-\frac{N}{Q_\infty} \left(-\log_2 N + \frac{1}{2} - \frac{\gamma}{\log 2} \right)$$

to Σ_1. For $j \ge 2$ there is only a *simple pole*, yielding a contribution

$$-\frac{N}{Q_\infty} \left[-\sum_{j\ge 2} a_j \frac{1}{2^{j-1} - 1} \right].$$

Now we turn to Σ_2:

$$\Sigma_2 = \frac{1}{Q_\infty} f_2(N)$$

with

$$f_2(x) = \sum_{j\ge1} a_j (j-2) \sum_{m\ge1} 2^{-m(j-2)} \psi_2\left(\frac{x}{2^m}\right).$$

We transform $f_2(x)$: (for $-2 < \Re s < -1$)

$$f_2^*(s) = \Gamma(s) \cdot \sum_{j\ge1} a_j (j-2) \frac{1}{2^{j-2-s}-1}.$$

The pole of interest is this time $s=-1$.

Again, $j=1$ must be treated separately, yielding a contribution

$$\frac{N}{Q_\infty}\left(-\log_2 N + \frac{1}{2} + \frac{1}{\log 2} - \frac{\gamma}{\log 2}\right)$$

to Σ_2. For $j \ge 2$ the contribution to Σ_2 is

$$\frac{N}{Q_\infty} \sum_{j\ge2} a_j (j-2) \frac{1}{2^{j-1}-1}.$$

Hence we have

$$\begin{aligned}\Sigma_1 + \Sigma_2 &\sim \frac{N}{Q_\infty}\left(\frac{1}{\log 2} + \sum_{j\ge2} a_j (j-1) \frac{1}{2^{j-1}-1}\right)\\ &= N R^*(-1).\end{aligned}$$

Therefore we obtain the final result

$$l_N \sim N\big(\alpha + 1 - R^*(-1)\big) = 0.37204\ldots \cdot N,$$

where, again, the *ubiquitous* fluctuating term of order N was not mentioned for brevity.

References

1. P.Flajolet, M.Régnier and R.Sedgewick, *Combinatorics on Words*, Springer NATO ASI Series F, Vol. 12, Berlin, 1985.
2. P.Flajolet and R.Sedgewick, *Digital Search Trees Revisited*, SIAM J. Computing **15** (1986), 748–767.
3. P.Kirschenhofer and H.Prodinger, *Einige Anwendungen der Modulfunktionen in der Informatik*, Sitzungsberichte der Österreichischen Akademie der Wissenschaften **197** (1988), 339–366.
4. D.E.Knuth, *The Art of Computer Programming, Vol.3*, Addison Wesley, Reading, Mass., 1973.

TU Vienna
Wiedner Hauptstrasse 8-10
A-1040 Vienna
Austria