

SOME FURTHER RESULTS ON
DIGITAL SEARCH TREES

Peter Kirschenhofer, Helmut Prodinger

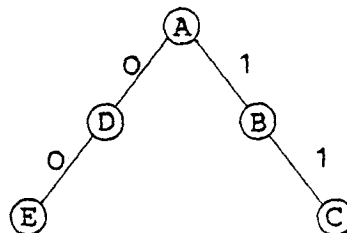
Institut für Algebra und Diskrete Mathematik
TU Vienna, A-1040 Vienna, Wiedner Hauptstr.8-10, Austria

1. INTRODUCTION

Our purpose in this paper is the study of digital search trees, (binary) tries and Patricia tries; a wealth of information about these important data structures can be found in Knuth's book [2]; compare also Flajolet and Sedgewick [1].

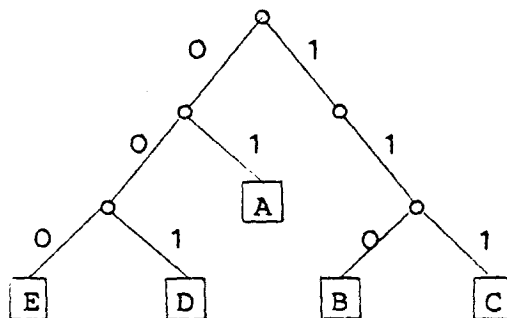
Considering digital search trees, we assume that each item has a key being an infinite sequence of 0 and 1, where 0 means "go left" and 1 means "go right", until an empty space is available for the insertion of the item:

- A : 010...
- B : 110...
- C : 111...
- D : 001...
- E : 000...

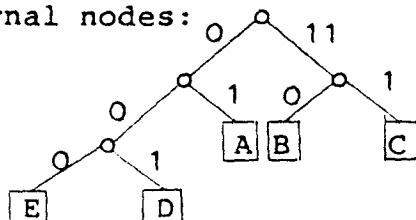


Note that the order in which the keys are inserted is relevant.

(Binary)tries follow the same idea, but the items are stored in the leaves, which makes the relative order of insertion irrelevant:



Patricia tries are constructed from tries by collapsing one-way branches on internal nodes:



In all 3 cases our parameter of interest is the number of nodes examined during a successful search, which corresponds to the internal path length in the digital search tree-case, and to the external path length in the two other cases.

The averages of these parameters were determined by Knuth [2] by means of the Mellin integral transform. Flajolet and Sedgewick gave alternative derivations in [1]; they use a rather simple but very useful formula due to S.O.Rice making the whole story easier and more transparent. (Compare Lemma 4.)

In Sections 2 to 4 of this paper we deal with the variances in all 3 cases; this quantity was never studied up to now. We use Rice's (or Flajolet's and Sedgewick's) method, since we feel that the original approach of Knuth might be too complicated and less transparent (even though we recently learned from W.Szpankowski that a Mellin transform approach might be feasible, compare [4],[5]).

In Section 2 we give a relatively detailed derivation for the instance of tries, since this case is the easiest one; in Sections 3 and 4 we continue with Patricia tries and digital search trees, respectively.

Since it is needed in the further considerations we cite the following result:

THEOREM 1[Knuth;Flajolet,Sedgewick]:

The expected value of the external path length of a trie resp. a Patricia trie built from N records with keys from random bit streams is

$$N(\log_2 N + \frac{\gamma}{\log 2} + \frac{1}{2} + \delta^{[T]}(\log_2 N)) + O(1)$$

resp.

$$N(\log_2 N + \frac{\gamma}{\log 2} - \frac{1}{2} + \delta^{[P]}(\log_2 N)) + O(1),$$

where γ is Euler's constant and $\delta^{[T]}(x) = \delta^{[P]}(x)$ is a periodic function with period 1 and very small amplitude:

$$\delta^{[T]}(x) = \frac{1}{\log 2} \sum_{k \in \mathbb{Z}, k \neq 0} \omega_k \cdot \Gamma(-\omega_k) e^{2k\pi i x}$$

with $\omega_k = 1 + 2k\pi i / \log 2$.

The expected value of the internal path length of a digital search tree built from N records with keys from random bit streams is

$$N(\log_2 N + \frac{\gamma-1}{\log 2} + \frac{1}{2} - \alpha + \delta^{[D]}(\log_2 N)) + O(N^{1/2})$$

where $\alpha = \sum_{k \geq 1} \frac{1}{2^k - 1} = 1,606695\dots$

and

$$\delta^{[D]}(x) = \frac{1}{\log 2} \sum_{k \in \mathbb{Z}, k \neq 0} \Gamma(-\omega_k) e^{2k\pi i x} . \square$$

So the averages are of order $N \cdot \log N$.

In sections 2 to 4 we will prove the following result on the variances which shows that they are of order N :

THEOREM 2: The variance is asymptotic to: In the case of

$$[T] \text{ (Tries)} \quad N \cdot \left(\frac{1}{12} + \frac{\pi^2}{6 \log^2 2} + \sigma^{[T]}(\log_2 N) \right)$$

[P] (Patricia tries)

$$N \cdot \left(\frac{1}{12} + \frac{\pi^2}{6 \log^2 2} - \frac{2}{\log 2} \log \prod_{\lambda \geq 1} \left(1 + \frac{1}{2^\lambda} \right) + \sigma^{[P]}(\log_2 N) \right)$$

[D] (Digital search trees)

$$N \cdot \left(\frac{1}{12} + \frac{\pi^2}{6 \log^2 2} + \frac{1}{\log^2 2} - \alpha - \beta + \sigma^{[D]}(\log_2 N) \right)$$

with α from Theorem 1, $\beta = \sum_{k \geq 1} \frac{1}{(2^k - 1)^2}$,

and the periodic functions

$$\sigma^{[T]}(x) = \frac{2}{\log^2 2} \sum_{k \neq 0} (\Gamma(-\omega_k) - \omega_k \Gamma'(-\omega_k) - \gamma \omega_k \Gamma(-\omega_k)) e^{2k\pi i x} - (\delta^{[T]}(x))^2,$$

$$\sigma^{[P]}(x) = \sigma^{[T]}(x) + \frac{2}{\log 2} \sum_{k \neq 0} \omega_k \Gamma(-\omega_k) (1 - \xi_k) e^{2k\pi i x}$$

$$\text{with } \xi_k = \sum_{\lambda \geq 1} \left[\left(1 + \frac{1}{2^\lambda} \right)^{\omega_k - 1} - 1 \right],$$

and

$$\sigma^{[D]}(x) = \frac{2}{\log^2 2} \sum_{k \neq 0} (-\Gamma'(-\omega_k) + (1 - \gamma) \Gamma(-\omega_k)) e^{2k\pi i x} - (\delta^{[D]}(x))^2 .$$

Ignoring the small fluctuations we have the following


COROLLARY 3: The variances are roughly

$$[T] \quad N \cdot 3,5070\dots$$

$$[P] \quad N \cdot 1,0000\dots$$

$$[D] \quad N \cdot 2,7841\dots$$

In the last section we consider the distribution of various types of nodes in the three types of data structures. This is a continuation of the investigations of [1], where Flajolet and Sedgewick have solved

an open problem of Knuth, namely to determine the expected number of internal nodes \odot of the type  in digital search trees built from N records.

$[z^k]f(z)$ denotes the coefficient of z^k in the power series $f(z)$.

2. TRIES

Let $h_N(z)$ be the generating function with $[z^k]h_N(z)$ the expected number of external nodes at level k in the family of tries built from N records with keys from random bit streams. Note that $h_N(1) = N$, the expected external path length is $h'_N(1)$ and the variance of this parameter (which will be computed now) is

$$h''_N(1) + h'_N(1) - \frac{1}{N}(h'_N(1))^2. \quad (2.1)$$

From [2] we know the recursion

$$h_N(z) = 2^{1-N} \sum_{k \geq 0} \binom{N}{k} z \cdot h_k(z), \quad N \geq 2, \quad (2.2)$$

$$h_0(z) = 0, \quad h_1(z) = 1.$$

We need some more generating functions:

$$R(z) = \sum_{N \geq 0} h'_N(1) \frac{z^N}{N!}, \quad S(z) = \sum_{N \geq 0} h''_N(1) \frac{z^N}{N!},$$

$$V(z) = e^{-z} \cdot R(z) = \sum_{N \geq 0} v_N \frac{z^N}{N!} \quad \text{and}$$

$$W(z) = e^{-z} \cdot S(z) = \sum_{N \geq 0} w_N \frac{z^N}{N!}.$$

From (2.2) we have

$$h''_N(1) = 2^{1-N} \sum_{k \geq 0} \binom{N}{k} (h''_k(1) + 2h'_k(1)), \quad N \geq 0,$$

and thus

$$S(z) = 2S\left(\frac{z}{2}\right) e^{z/2} + 4R\left(\frac{z}{2}\right) e^{z/2}$$

resp.

$$W(z) = 2W\left(\frac{z}{2}\right) + 4V\left(\frac{z}{2}\right).$$

From [1] we have

$$v_N = \frac{N(-1)^N}{1-2^{1-N}}, \quad N \geq 2,$$

so that

$$w_N = \frac{2 \cdot 2^{1-N} \cdot N \cdot (-1)^N}{(1-2^{1-N})^2}, \quad N \geq 2; \quad w_0 = w_1 = 0.$$

By $S(z) = e^z \cdot W(z)$ we get

$$h_N''(1) = \sum_{k \geq 2} \binom{N}{k} (-1)^k \frac{2 \cdot 2^{1-k} \cdot k}{(1-2^{1-k})^2} \quad (2.3)$$

The asymptotic evaluation of this alternating sum is now attacked by "Rice's method" using a classical formula for finite differences [3].

LEMMA 4[Nörlund]: Let C be a curve surrounding the points $2, \dots, N$ and $f(z)$ be analytic within C . Then

$$\sum_{k \geq 2} \binom{N}{k} (-1)^k f(k) = \frac{-1}{2\pi i} \int_C [N; z] f(z) dz$$

$$\text{with } [N; z] = \frac{(-1)^{N-1} N!}{z(z-1)\dots(z-N)}.$$

Applying the Lemma to expression (2.3) and moving the contour of integration to the left of the line with $\text{Re } z = 1$ (compare [1] for technical details), we obtain

$$h_N''(1) \sim \sum_{k \in \mathbb{Z}} \text{Res}([N; z] f(z); z = \omega_k) \quad (2.4)$$

In order to determine the residues, we have to work out the local expansions of

$$[N; z] f(z) = [N; z] \cdot \frac{2 \cdot 2^{1-z} \cdot z}{(1-2^{1-z})^2} \quad (2.5)$$

about the poles $\omega_0 = 1$ (triple) and $\omega_k, k \neq 0$, (double). For this purpose we manage to get a list of local expansions of the "ingredients":

With $u = z-1$ and $L = \log 2$ we have:

$$[N; z] \sim \frac{N}{u} \left(1 + u(H_{N-1} - 1) + u \left(1 - H_{N-1} + \frac{1}{2} H_{N-1}^2 + \frac{1}{2} H_{N-1}^{(2)} \right) \right)$$

(H_N resp. $H_N^{(2)}$ denoting Harmonic numbers)

$$\begin{aligned} 2^{1-z} &\sim 1 - Lu + \frac{L^2 u^2}{2} \\ \frac{1}{1-2^{1-z}} &\sim \frac{1}{Lu} \left(1 + \frac{Lu}{2} + \frac{L^2 u^2}{12} \right) \\ \frac{1}{(1-2^{1-z})^2} &\sim \frac{1}{L^2 u^2} \left(1 + Lu + \frac{5}{12} L^2 u^2 \right). \end{aligned} \quad (2.6)$$

The expansions in $u = z - \omega_k$ read:

$$\begin{aligned} [N; z] &\sim N^{\omega_k} [\Gamma(-\omega_k) + u(-\Gamma'(-\omega_k) + \Gamma(-\omega_k) \log N)] \\ 2^{1-z} &\sim 1 - Lu \end{aligned}$$

$$\frac{1}{1-2^{1-z}} \sim \frac{1}{Lu} \left(1 + \frac{Lu}{2}\right) \quad (2.7)$$

$$\frac{1}{(1-2^{1-z})^2} \sim \frac{1}{L^2 u^2} (1 + Lu) .$$

Calculating the residues of (2.5) we obtain by (2.4)

$$h_N''(1) \sim \frac{N}{L^2} (H_{N-1}^2 + H_{N-1}^{(2)}) - \frac{N}{6} + \\ + \frac{2N}{L^2} \sum_{k \neq 0} N^{\omega_k - 1} (\Gamma(-\omega_k) + \omega_k (-\Gamma'(-\omega_k) + \Gamma(-\omega_k) \log N)).$$

Inserting the asymptotic expansion of the Harmonic numbers and using the asymptotic equivalents for $h_N'(1)$ from Theorem 1, we get Theorem 2 [T].

3. PATRICIA TRIES

We keep the notation of Section 2, but now always referring to Patricia tries.

The recurrence relation for $h_N(z)$ is now

$$h_N(z) = 2^{1-N} \sum_{k \geq 0} \binom{N}{k} z h_k(z) - 2^{1-N} (z-1) h_N(z), \quad N \geq 2; \quad h_0(z) = h_1(z) = 1. \quad (3.1)$$

The functional equation for $W(z)$ reads

$$W(z) = 2W\left(\frac{z}{2}\right) + 4(1-e^{-z/2})V\left(\frac{z}{2}\right) \quad (3.2)$$

where ([1])

$$V_N = \frac{N(-1)^N}{2^{N-1}-1}, \quad N \geq 2.$$

Hence

$$h_N''(1) = \sum_{k \geq 2} \binom{N}{k} \frac{2k(-1)^k}{(2^{k-1}-1)^2} - \sum_{k \geq 2} \binom{N}{k} \frac{2k(-1)^k}{2^{k-1}-1} \sum_{\lambda \geq 1} \left[\left(1 + \frac{1}{2^\lambda}\right)^{k-1} - 1 \right]. \quad (3.3)$$

So Rice's method can be used with

$$f(z) = \frac{2z}{(2^{z-1}-1)^2} - \frac{2z}{2^{z-1}-1} \sum_{\lambda \geq 1} \left[\left(1 + \frac{1}{2^\lambda}\right)^{z-1} - 1 \right]. \quad (3.4)$$

We may now use the local expansions from (2.6) and (2.7) together with $(z-1)$:

$$\sum_{\lambda \geq 1} \left[\left(1 + \frac{1}{2^\lambda}\right)^{z-1} - 1 \right] \sim (z-1) \cdot \log \prod_{\lambda \geq 1} \left(1 + \frac{1}{2^\lambda}\right)$$

and Theorem 2 [P] follows immediately.

4. DIGITAL SEARCH TREES

From Knuth [2, p.496] we have ($[z^k]h_N(z)$ now referring to internal nodes)

$$h_N(z) = \sum_{k \geq 0} \binom{N}{k+1} (-1)^k \prod_{0 \leq j < k} \left(1 - \frac{z}{2^j}\right) \quad (4.1)$$

so that, after some easy manipulations,

$$h_N''(1) = 2 \sum_{k \geq 2} \binom{N}{k} (-1)^{k-1} Q_{k-2} \cdot T(k-2) \quad (4.2)$$

with (compare [1])

$$Q_N = Q(1)/Q(2^{-N}) \text{ and } Q(z) = \prod_{j \geq 1} \left(1 - \frac{z}{2^j}\right)$$

and

$$T(k) = \sum_{1 \leq j \leq k} \frac{1}{2^{j-1}}.$$

The appropriate extension of $T(k)$ to \mathbb{C} is

$$T(z) = \alpha - \sum_{j \geq 1} \frac{1}{2^{z+j-1}}.$$

Rice's method can be applied, taking the local expansions for $Q(z)$ from [1] extended by one more term; the local expansion of $T(z-2)$ is

$$T(z-2) \sim -\frac{1}{L} \frac{1}{z-1} + \frac{1}{2} + (z-1)L \left[-\frac{1}{12} + \alpha + \beta \right]$$


and

$$T(z-2) \sim -\frac{1}{L} \frac{1}{z-\alpha_k} + \frac{1}{2}.$$

Calculating the residues from the local expansions, which is straightforward but tedious, we arrive at part [D] of Theorem 2.

5. DISTRIBUTION OF VARIOUS TYPES OF NODES

In solving an open problem due to Knuth, Flajolet and Sedgewick [1] have counted the average number $A_N^{[D]}$ of nodes with both sons external

nodes  in digital search trees built from N records with keys

from random bit streams:

THEOREM 5 [1]:

$$A_N^{[D]} \sim N \cdot (\mu + \tau^{[D]}(\log_2 N))$$



with

$$\mu = \sum_{k \geq 1} \frac{k 2^{k+1}}{1 \cdot 3 \cdots (2^k - 1)} \sum_{1 \leq j \leq k} \frac{1}{2^{j-1}} + 1 - \left(\prod_{k \geq 1} \frac{2^k}{2^k - 1} \right) \cdot \left(\frac{1}{\log 2} + \alpha^2 - \alpha \right)$$

and

$$\tau^{[D]}(x) = \frac{1}{\log 2} \left(\prod_{k \geq 1} \frac{2^k}{2^{k-1}} \right) \sum_{k \neq 0} (\omega_k - 1) \Gamma(\omega_k - 2) e^{2k\pi i x}. \square$$

The corresponding averages $B_N^{[D]}$ and $C_N^{[D]}$ of internal nodes \odot of

type  resp.  are then related to $A_N^{[D]}$ by the relations

$$2A_N^{[D]} + 2B_N^{[D]} = N+1 \quad (\text{enumerating leaves})$$

and

$$A_N^{[D]} + 2B_N^{[D]} + C_N^{[D]} = N \quad (\text{enumerating internal nodes})$$

so that

COROLLARY 6: $B_N^{[D]} \sim \frac{N}{2} \cdot (1 - \mu - \tau^{[D]}(\log_2 N))$

$$C_N^{[D]} \sim A_N^{[D]}.$$

Now we turn our attention to tries built from N records and the

averages $A_N^{[T]}, B_N^{[T]}, C_N^{[T]}$ and $D_N^{[T]}$ of internal nodes \odot of type



The average ℓ_N of the total number of internal nodes is, implicitly, given in Knuth [2, p.494]:

$$\ell_N \sim \frac{N}{\log 2} \left[1 + \sum_{k \neq 0} \Gamma(-\omega_k) (\omega_k - 1) e^{2k\pi i \cdot \log_2 N} \right] \quad (5.1)$$

We have the following relations

$$2A_N^{[T]} + 2B_N^{[T]} = N \quad (\text{enumerating leaves})$$

$$2A_N^{[T]} + 2B_N^{[T]} + 2D_N^{[T]} = \ell_N + 1 \quad (\text{enumerating leaves of the extended binary tree})$$

$$A_N^{[T]} + 2B_N^{[T]} + C_N^{[T]} + 2D_N^{[T]} = \ell_N \quad (\text{enumerating internal nodes}).$$

Thus we have

$$B_N^{[T]} = \frac{N}{2} - A_N^{[T]}$$

$$C_N^{[T]} = A_N^{[T]} - 1 \quad (5.2)$$

$$D_N^{[T]} = \frac{1}{2} (\ell_N + 1 - N).$$

For $A_N^{[T]}$ we have the recurrence relation

$$A_N^{[T]} = \sum_{k \geq 0} \frac{1}{2^N} \binom{N}{k} (A_k^{[T]} + A_{N-k}^{[T]}), N \geq 3, A_0^{[T]} = A_1^{[T]} = 0, A_2^{[T]} = 1. \quad (5.3)$$

Using generating functions as before

$$A_N^{[T]} = \sum_{k \geq 2} \binom{N}{k} \frac{(-1)^k k(k-1)}{4(1-2^{1-k})}$$

and Rice's method can be applied to get

THEOREM 7:

$$A_N^{[T]} \sim \frac{N}{4 \log 2} \left(1 + \sum_{k \neq 0} \omega_k (\omega_k - 1) \Gamma(-\omega_k) e^{2k\pi i \cdot \log_2 N} \right).$$

The corresponding averages for Patricia tries are $A_N^{[P]} = A_N^{[T]}$,
 $B_N^{[P]} = B_N^{[T]}$ and $C_N^{[P]} = C_N^{[T]}$, ($D_N^{[P]} = 0!$) because of their construction
 from tries.

REFERENCES

- [1] P.Flajolet, R.Sedgewick, Digital Search Trees Revisited, SIAM Journal on Computing, to appear 1986.
- [2] D.E.Knuth, The Art of Computer Programming, Vol.3: Sorting and Searching, Addison-Wesley, Reading Mass. 1973.
- [3] N.E.Nörlund, Vorlesungen über Differenzenrechnung, Chelsea, New York 1954.
- [4] W.Szpankowski, Analysis of a recurrence equation arising in stack-type algorithms for collision-detecting channels, Proc.of Intern.Symp.on Computer Networking & Performance Evaluation, Sept.1985, Tokyo.
- [5] W.Szpankowski, Solution of linear recurrence equations arising in analysis of some algorithms, preprint.