INFORMATION PROCESSING 89, G.X. Ritter (ed.)
Elsevier Science Publishers B.V. (North-Holland)

DIGITAL SEARCH TREES - FURTHER RESULTS ON A FUNDAMENTAL DATA STRUCTURE

Peter KIRSCHENHOFER and
Helmut PRODINGER

Technical University of Vienna
Dept. of Algebra and Discrete Mathematics
A-1040 Vienna, Austria

Wojciech SZPANKOWSKI

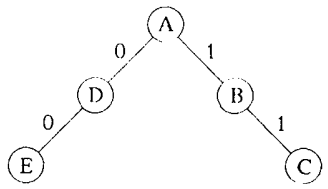
Purdue University
Dept. of Computer Science
West Lafayette, IN 47907, U.S.A.

This paper deals with the average case performance of a prominent data structure, namely digital search trees which are useful in many applications in Computer Science and Telecommunication, such as partial match retrieval of multidimensional data, conflict resolution algorithms for broadcast communication, radix exchange sort, polynomial factorization, simulation, lexicographical sorting and extendible hashing.
The most important parameter related to the cost of successful search is the so called *internal path length*. Knuth, Flajolet and others gave asymptotic results on the expectation of this quantity, whereas the analysis of the variance was open up to now. In this paper we solve this problem. The solution relies on deep analytic tools which we sketch in the final section. It turns out that while (apart from some ubiquitous small periodic fluctuations) the expectation is of order $N \log N$, the variance is only of order N. (N refers to the size of the data structure.)
This analysis completes previous results on Tries and Patricia Tries.

1.INTRODUCTION

Analysis of algorithms is a rapidly developing area in Theoretical Computer Science with a strong impact to practical problems. The IFIP congress has a long tradition in presenting results in that area which has been started with the famous paper *"Mathematical Analysis of Algorithms"* by *Donald Knuth* in 1971 [11]. The cost of the performance of algorithms, including topics as the storage requirements of data structures and the execution time of certain subroutines, is usually described in terms of worst case behaviour and average case behaviour. Sophisticated methods have been applied in order to optimize the worst case behaviour of algorithms; nevertheless for practical purposes the question of the *average case performance of data structures and algorithms* is considered more and more to be of great importance.
Digital tree search was first proposed by Coffman and Eve in 1970 [3]: the digital search tree is a data structure which leads to asymptotically optimal average case performance by using the digital properties of keys. We assume that the keys are represented as sequences of bits. The keys are inserted into the nodes of a binary tree exactly as with binary search trees. However, the left/right decision is governed by the bits of the keys. For example

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



Note that the order in which the keys are inserted is relevant.
We mention for later comparisions that so called *tries* and *Patricia tries* follow basically the same idea but storing keys in leaves despite of internal nodes, compare [12].
Digital search trees and tries find many applications - in Computer Science and Telecommunication such as partial match retrieval for multidimensional data, conflict resolution algorithms for broadcast communication [14], radix exchange sort, polynomial factorization, simulation [6], [12], lexicographical sorting [1], [16] and extendible hashing [4].
The average number of nodes examined during a successful search in a search tree with N nodes is 1 + the *internal path length*, divided by N:
The internal path length is the sum of the lengths of the paths from the root to each node in the tree (in our example it is 6).
Let l_N be the *expected internal path length* of a digital tree built from N (sufficiently long keys) comprised of random bits. Then the following asymp-

totic result holds:

$$l_N \sim N \cdot \log_2 N + N \cdot (-1.716... + \delta_1(\log_2 N)), \qquad (1)$$

where $\delta_1(x)$ is a periodic function of very small amplitude ($<10^{-6}$). So for practical purposes δ_1 may be neglected as far as the expectation is concerned. Nevertheless it turns out that a careful analysis of δ_1 (as well as the smaller order terms of l_N) is of crucial importance for the analysis of the *variance*. (1) has been established by Konheim-Newman [13], Knuth [12], Flajolet-Sedgewick [5], using different technical tools. However, the analysis of the variance has turned out to be an even more challenging problem and was open up to now; this paper gives an outline to the final answer to this question including remarks on the difficult technical *apparatus* that is needed for the achievement of this result.
(The analysis of the (external) path length in tries resp. Patricia tries is remarkably easier but by no means trivial, compare [9] and [10]. The same parameters in an easier statistical model, namely considering *abstract averaging trees*, were studied in [7] and [17].)
We conclude this section with a table of expectations and variances of the internal resp. external path lengths (ignoring the ubiquitous small periodic fluctuations):

	Expectation	Variance
Digital search trees	$N(\log_2 N - 1.71)$	$N \cdot 0.25$
Tries	$N(\log_2 N + 1.33)$	$N \cdot 4.35$
Patricia tries	$N(\log_2 N + 0.33)$	$N \cdot 0.37$

It can be seen from this table that the variance is rather small (of order N (!)); again digital search trees prove to be the best incarnation of the idea of digital searching.

2. OUTLINE OF THE METHODS AND MAIN RESULTS

Let T_N be the family of digital search trees built from N records with keys from random bit streams. A key consists of 0's and 1's with equal probability of appearance. Let X_N denote the random variable "internal path length" of trees in T_N and $F_N(z)$ the corresponding probability generating function, i.e. the coefficient $[z^k]\, F_N(z)$ of z^k in $F_N(z)$ is the probability that a tree in T_N has internal path length equal to k. Then the following recursion holds which is immediate from the definitions:

$$F_{N+1}(z) = z^N \sum_{k=0}^{N} 2^{-N} \binom{N}{k} F_k(z)\, F_{N-k}(z) \qquad (N\geq 0),\ F_0(z) = 1\ . \qquad (2)$$

Then the *expectation* l_N is given by $l_N = F_N'(1)$ and fulfills

$$l_{N+1} = N + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k \qquad (N>0),\ l_0 = 0 \qquad (3)$$

This recursion may be solved explicitely by the use of *exponential generating functions:*

With $L(z) := \sum_{N\geq 0} l_N \frac{z^N}{N!}$ (3) translates into the following functional differential equation

$$L'(z) = ze^z + 2\, e^{z/2} L(\tfrac{z}{2}) \qquad (4)$$

By the substitution $\hat{L}(z) = e^z L(-z)$ we have the easier equation

$$\hat{L}(z) - \hat{L}'(z) = -z + 2\, \hat{L}(\tfrac{z}{2})\ . \qquad (5)$$

With $\hat{L}(z) = \sum_{N\geq 0} \hat{l}_N \frac{z^N}{N!}$ we find

$$\hat{l}_N = Q_{N-2} \quad (N\geq 2), \qquad \hat{l}_0 = \hat{l}_1 = 0 \qquad (6)$$

with the finite product

$$Q_N = (1 - \tfrac{1}{2})(1 - \tfrac{1}{4}) \cdots (1 - \tfrac{1}{2^N}) \qquad (7)$$

so that finally

$$l_N = \sum_{k=2}^{N} \binom{N}{k} (-1)^k\, Q_{k-2}\ . \qquad (8)$$

The reader should note that an asymptotic evaluation of (8) is non elementary by the fact that terms of almost equal magnitude occur with alternating signs. For this reason sophisticated methods from complex analysis are needed to find the correct order of growth. An essential step is the application of the following lemma from the calculus of finite differences.

Lemma 1. (Compare [12;p.138], [15]). Let $\mathfrak{C}$ be a path surrounding the points j,j+1,...,N and f(z) be analytic inside $\mathfrak{C}$. Then

$$\sum_{k\geq j} \binom{N}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_{\mathfrak{C}} [N;z]\, f(z)\, dz \qquad (9)$$

with the abbreviation $[N;z] = \frac{(-1)^{N-1}\, N!}{z(z-1)\cdots(z-N)}$. □

In our applications f(z) is a meromorphic function that continues a sequence f(k), e.g. j=2 and $f(k) = Q_{k-2}$ in (8). Moving the contour of integration it turns out that the asymptotic expansion of the alternating sum is obtained via

$$\Sigma\ \mathrm{Res}\ \big([N;z]\, f(z)\big)\ ,$$

where the sum is taken over all poles different from j,j+1,...,N.

For technical details of the continuation of $f(k) = Q_{k-2}$ we refer to the next section. Regarding the residues at poles with real part > -1 we derive the following accurate asymptotic expansion for l_N.

Theorem 2. The expectation l_N of the internal path length of digital search trees built from N records fulfills

$$l_N \sim N \log_2 N + N\Big(\frac{\gamma-1}{\log 2} + \frac{1}{2} - \alpha + \delta_1(\log_2 N)\Big)$$
$$+ \log_2 N + \frac{2\gamma-1}{2\log 2} + \frac{5}{2} - \alpha + \delta_2(\log_2 N) \qquad (10)$$

with $\gamma = 0.57721\cdots$ (Euler's constant) and $\alpha = \sum_{n\geq 1} 1/(2^n - 1) = 1.60669$. $\delta_1(x)$ and $\delta_2(x)$ are continuous periodic functions of period 1, mean 0 and very small amplitude; for later use we mention the Fourier expansion of $\delta_1(x)$:

$$\delta_1(x) = \frac{1}{\log 2} \sum_{k\neq 0} \Gamma\Big(-1 - \frac{2k\pi i}{\log 2}\Big) e^{2k\pi i x}. \quad \square \qquad (11)$$

We mention in passing that the $O(1)$-term in (10) is slightly incorrect in [12].

Now we turn to the analysis of the variance which is given by

$$\mathrm{Var}\, X_N = s_N + l_N - l_N^2 \qquad (12)$$

with $s_N = F_N''(1)$. From (2) we get the recurrence relation (for $N\geq 0$; $s_0=0$)

$$s_{N+1} = N 2^{2-N} \sum_{k=0}^{N} \binom{N}{k} l_k + N(N-1)$$
$$+ 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k l_{N-k} + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} s_k \,. \qquad (13)$$

In order to find an explicit solution to this recurrence it is split up into 3 parts: $s_N = u_N + v_N + w_N$, where

$$u_N = 2N(l_{N+1} - N) + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} u_k \,,$$
$$N\geq 0,\ u_0 = 0 \qquad (14)$$

$$v_N = N(N-1) + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} v_k, \quad N\geq 0,\ v_0 = 0 \qquad (15)$$

$$w_N = 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k l_{N-k} + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} w_k \,,$$
$$N\geq 0,\ w_0 = 0 \qquad (16)$$

A similar treatment as with (3) leads to the following explicit solutions

$$u_N = \sum_{k=3}^{N} \binom{N}{k} (-1)^k \hat{u}_k \quad , u_0 = u_1 = u_2 = 0$$

where (17)

$$\hat{u}_k = 2Q_{k-2}\Big[4 + \sum_{j=1}^{k-2} \frac{1}{2^{j-1}} - \sum_{j=1}^{k-2} \frac{j}{2^{j-1}} - \frac{2k}{2^{k-2}-1}\Big] \,,$$

$$v_N = \sum_{k=3}^{N} \binom{N}{k} (-1)^k \hat{v}_k \,, \quad v_0 = v_1 = v_2 = 0$$

where (18)

$$\hat{v}_k = -4\, Q_{k-2} \,,$$

$$w_N = \sum_{k=5}^{N} \binom{N}{k} (-1)^k \hat{w}_k \,, \quad w_0 = \ldots = w_4 = 0$$

where (19)

$$\hat{w}_k = -\sum_{j=4}^{k-1} 2^{1-j} \sum_{i=2}^{j-2} \binom{j}{i} \frac{Q_{i-2}\, Q_{j-i-2}\, Q_{k-2}}{Q_{j-1}} \,.$$

It is not too difficult to prove that

$$v_N = 4\binom{N}{2} - 4\, l_N, \qquad (20)$$

so that the treatment of u_N and w_N remains to be done.

In principle u_N and w_N may be analyzed by making use of Lemma 1. However it turns out to be a highly non trivial problem to find continuations of $\hat{u}_k$ resp. $\hat{w}_k$. We refer to that problem in the next section. After lengthy and difficult computations the residue calculus leads us to the following results.

Lemma 3. With the abbreviation $L = \log 2$ we have

$$u_N \sim \frac{4}{L} N^2 \log N$$
$$+ N^2\Big(\frac{4(\gamma-1)}{L} - 6 - 4\alpha + \delta_3(\log_2 N)\Big)$$
$$- \frac{N \log^2 N}{L^2} + \frac{2N}{L} \log N \cdot \Big(\frac{2-\gamma}{L} + 8 + \alpha\Big)$$
$$+ N\big(c_1 + \delta_4(\log_2 N)\big) \qquad (21)$$

$$w_N \sim \frac{1}{L^2} N^2 \log^2 N \qquad (22)$$
$$+ N^2 \log N \cdot \Big(-\frac{3}{L} - \frac{2}{L^2} + \frac{2\gamma}{L^2} - \frac{2\alpha}{L}\Big)$$
$$+ N^2 \cdot \Big(\frac{14}{3} Q_\infty^2 + \frac{1}{3} + \alpha^2 + 3\alpha + \frac{2\alpha}{L} - \frac{3\gamma}{L}$$
$$- \frac{2\alpha\gamma}{L} - \frac{\beta_1}{L} + \frac{2}{L} + \frac{2}{L^2} + \frac{\gamma^2}{L^2}$$
$$+ \frac{\pi^2}{6L^2} - \frac{2\gamma}{L^2} + \delta_5(\log_2 N)\Big)$$
$$+ \frac{3}{L^2} N \log^2 N$$
$$+ N \log N \cdot \Big(-\frac{7}{L^2} - \frac{3}{L} - \frac{6\alpha}{L} - \frac{10\gamma}{L^2}\Big)$$
$$+ N\big(c_2 + \delta_6(\log_2 N)\big)$$

with some very involved constants c_1, c_2 and periodic functions $\delta_3(x)$,..., $\delta_6(x)$ of mean zero and

$$\beta_1 = 2\sum_{k\ge 2}\frac{(-1)^k}{(k+1)k(k-1)(2^k-1)} \quad . \ \square \tag{23}$$

For completeness we mention the asymptotics for v_N, derived from (20):

$$v_N \sim 2N^2 - \frac{4}{L}N\log N + N\left(c_3+\delta_7(\log_2 N)\right) \tag{24}$$

According to (12) we thus find easily that the terms of order $N^2\log^2 N$ and $N^2\log N$ cancel in Var X_N. However, the cancellation of the terms of order N^2 is non trivial, since it contains the square δ_1^2 of the periodic fluctuation originating from h_N. From a deep result in the theory of modular functions it can be shown that the mean $[\delta_1^2]_0$ of δ_1^2 is given by (compare [8]):

Lemma 4.
$$[\delta_1^2]_0 = \frac{1}{L^2}\sum_{k\neq 0}\left|\Gamma\left(-1-\frac{2k\pi i}{\log 2}\right)\right|^2$$

$$= \frac{1}{L^2} + \frac{\pi^2}{6L^2} - \frac{1}{L} - \frac{\beta_1}{L} - \frac{47}{12} \quad . \qquad \square$$

Thus we have

$$\text{Var } X_N \sim N^2\,\delta_8(\log_2 N), \tag{25}$$

where $\delta_8(x)$ has mean 0. From a continuity argument (see Section 3) it follows that $\delta_8(x) \equiv 0$. Similarily we can prove that the terms of order $N\log^2 N$ and $N\log N$ cancel, so that

$$\text{Var } X_N \sim N\left(c_4 + \delta_9(\log_2 N)\right).$$

Theorem 5. The variance of the internal path length of digital search trees built from N records fulfills

$$\text{Var } X_N \sim N\left(0.25 + \delta_9(\log_2 N)\right),$$

where $\delta_9(x)$ is a continuous function of period 1, mean 0 and very small amplitude.

3. SOME TECHNICAL REMARKS

In the sequel we stress some of the analytic problems that we encounter during the investigation. We first mention the problem of finding appropriate analytic continuations for certain discrete sequences a needed for the application of Lemma 1. The first instance is

$$f(k) = Q_{k-2} = \prod_{i=1}^{k-2}\left(1-\frac{1}{2^i}\right) \quad .$$

This can be handled by considering the infinite product

$$Q(x) = \prod_{k\ge 1}\left(1-\frac{x}{2^k}\right) \quad , \tag{26}$$

so that

$$Q_{k-2} = Q_\infty / Q(2^{2-k})$$

with

$$Q_\infty = Q(1) = \prod_{k\ge 1}\left(1-\frac{1}{2^k}\right) = 0.288\cdots \quad . \tag{27}$$

Thus we may take

$$f(z) = Q_\infty / Q(2^{2-z})$$

which is a meromorphic function with poles in

$$z = m + \frac{2k\pi i}{\log 2}, \quad m = 1, 0, -1, -2,\ldots; \ k\in\mathbb{Z}.$$

It turns out that the poles with non-zero imaginary part constitute the periodic fluctuations in the asymptotic expansions - a phenomenon that occurs frequently in Computer Science [12].

The most challenging problem in this context occurs with the continuation of $\hat{w}_k$: After some manipulations starting from (19) we may write $\hat{w}_k$ as

$$\hat{w}_k = Q_{k-2}\cdot\left[-2(k-4)\,Q_\infty - 2Q_\infty^2\,B(k) - \tau + T(k)\right]$$

with

$$B(k) = \sum_{j=3}^{k-2}\left[\frac{1}{Q_j} - \frac{1}{Q_\infty}\right]$$

and

$$T(k) = \sum_{j\ge 0}\sigma(j+k+1), \qquad \tau = T(4),$$

where

$$\sigma(n) = -\frac{2Q_\infty^2}{Q_{n-2}} + \frac{\xi(n)}{2^{n-2}\,Q_{n-2}}$$

and

$$\xi(n) = \sum_{k=2}^{n-3}\binom{n-1}{k}Q_{k-2}\,Q_{n-3-k}.$$

To continue B(k) is not too difficult; we may take

$$B(z) = \sum_{j\ge 3}\left[\frac{1}{Q_j} - \frac{1}{Q_\infty}\right] - \sum_{j\ge 1}\left[\frac{1}{Q_{z-2+j}} - \frac{1}{Q_\infty}\right]$$

(where $Q_z = Q_\infty/Q(2^{-z})$ as before).

However, the continuation of ξ is not at all obvious because of the convolution involved. The following idea proves to be essential:

We consider the partial fraction decomposition

$$\frac{Q_\infty}{Q(x)} = \sum_{i\ge 1}\frac{a_i}{1-x2^{-i}}, \quad \text{with } a_i = (-1)^{i-1}/2^{\binom{i}{2}}Q_{i-1}.$$

Using this for the Q's occuring in the convolution and also some symmetry properties of the appearing terms we are finally led to

$$\xi(z+1) = \sum_{i,j\ge 1}a_ia_j\Bigg\{\frac{2^z-2-2z}{1-2^{-i-j-z+4}}$$

$$+\frac{1}{1-2^{-i-j-z+4}}\Bigg[\sum_{k>2}\binom{z}{k}\left(\frac{1}{2^{i+k-2}-1}+\frac{1}{2^{j+k-2}-1}\right)$$

$$-z\left(\frac{1}{2^{i+z-3}-1}+\frac{1}{2^{j+z-3}-1}\right)$$

$$-\left(\frac{1}{2^{i+z-2}-1}+\frac{1}{2^{j+z-2}-1}\right)\Bigg]\Bigg\}.$$

Next we mention that Lemma 4 follows from some identities related to *Dedekind's transformation formula* for the η-function (see [2])

$$\eta(\tau) = e^{\pi i \tau/12} \prod_{n\geq 1} \left(1 - e^{2\pi i n \tau}\right), \quad \Im(\tau) > 0.$$

Finally we turn our attention to the disappearing fluctuation in (25): $\delta_8(x)$ is a continuous periodic function of period 1 of mean 0, since its Fourier series is absolutely convergent. If $\delta_8(x)$ would not vanish identically we could find an $\varepsilon>0$ and an interval, say $[a,b]\subseteq[0,1]$, such that $\delta_8(x) < -\varepsilon$ for $x\in[a,b]$. Since $\log_2 N$ is dense modulo 1, the variance $\mathrm{Var}\, X_N$ would be negative for an infinity of values, an obvious contradiction.

REFERENCES

[1] A. Aho, J. Hopcroft, J. Ullman, Data Structures and Algorithms, Addison Wesley, Reading MA, 1983.

[2] T. M. Apostol, Modular Functions and Dirichlet Series in Number Theory, Springer, New York 1976.

[3] E. G. Coffman Jr., J. Eve, File structures using hashing functions, Comm. ACM **13** (1970), 427-436.

[4] R. Fagin, J, Nievergelt, N, Pippenger, H. Strong, Extendible hashing: A fast access method for dynamic files, ACM TODS **4** (1979), 315-344.

[5] P. Flajolet, R. Sedgewick, Digital Search Trees Revisited, SIAM J. Comput. **15** (1986), 748-767.

[6] G. Gonnet, Handbook of Algorithms and Data Structures, Addison Wesley, Reading MA, 1983.

[7] P. Kirschenhofer, H. Prodinger, Further results on digital search trees, Theor. Comput. Sci. **58** (1988), 143-154.

[8] P. Kirschenhofer, H. Prodinger, On some applications of formulae of Ramanujan in the Analysis of Algorithms, preprint 1987.

[9] P. Kirschenhofer, H. Prodinger, W. Szpankowski, On the variance of the external path length in a binary digital trie, Discrete Appl. Math., to appear.

[10] P. Kirschenhofer, H. Prodinger, W. Szpankowski, On the balance property of Patricia tries: External path length viewpoint, Theoret. Comput. Sci., to appear.

[11] D. E. Knuth, Mathematical analysis of algorithms, Information Processing 71, 19-27, North-Holland, Amsterdam 1972.

[12] D. E. Knuth, The Art of Computer Programming, Vol.3, Addison-Wesley,, Reading MA, 1973.

[13] A. G. Konheim, D. J. Newman, A note on growing binary trees, Discrete Mathematics, **4** (1973), 57-63.

[14] P. Mathys, P. Flajolet, Q-ary collision resolution algorithms in random-access systems with free and blocked channel access, IEEE IT **31**, 217-243 (1985).

[15] N. E. Nörlund, Vorlesungen über Differenzenrechnung, Chelsea, New York 1954.

[16] R. Paige, R. Tarjan, Three efficient algorithms based on partition refinement, preprint 1986.

[17] W. Szpankowski, Some results on v-ary asymmetric tries, J. Algorithms **9** (1988), 224-244.