Computing 28, 367—371 (1982)

Computing

A Note on the Average Depth of Tries

L. Devroye, Montreal

Received September 16, 1981

Abstract — Zusammenfassung

A Note on the Average Depth of Tries. Let A_n be the average root-to-leaf distance in a binary trie formed by the binary fractional expansions of n independent random variables $X_1, \ldots, X_n$ with common density f on $[0, 1)$. We show that *either* $E(A_n) = \infty$ for all $n \geq 2$ *or* $\lim_n E(A_n)/\log_2 n = 1$ depending on whether $\int f^2(x)\,dx = \infty$ or $\int f^2(x)\,dx < \infty$.

Key words and phrases: Trie, average depth, binary tree, random tree, successful search time.

Eine Bemerkung über die mittlere Höhe von Bäumen. Sei A_n der mittlere Wurzel-zu-Blatt-Abstand in einem binären Baum, der durch die Dualbruchentwicklungen von n unabhängigen Zufallsveränderlichen $X_1, \ldots, X_n$ mit gemeinsamer Dichte f auf $[0, 1)$ entsteht. Wir zeigen, daß *entweder* $E(A_n) = \infty$ für alle $n \geq 2$ *oder* $\lim_n E(A_n)/\log_2 n = 1$, je nachdem, ob $\int f^2(x)\,dx = \infty$ oder $\int f^2(x)\,dx < \infty$.

1. Introduction

A *trie* is a kind of binary search tree originally introduced by Fredkin (1960). We are given n countable strings of 0's and 1's, say $X_1, \ldots, X_n$, and we consider the infinite binary tree formed by the paths that correspond to the X_i's ("0" stands for a left turn down the tree, and "1" indicates a right turn). The trie formed by $X_1, \ldots, X_n$ is the smallest subtree of this tree with the property that all n truncated paths are pairwise different. The X_i's are then associated with the n leaves of this binary tree. For a fairly comprehensive treatment of tries, with applications, see Knuth (1973b).

Let D_{ni} be the depth of X_i (distance from the root) in the trie formed by $X_1, \ldots, X_n$. The average successful search time for the given trie is equal to the average depth:

$$A_n = \frac{1}{n} \sum_{i=1}^{n} D_{ni}.$$

We would like to say something meaningful about the average depth of a trie, and it is clear that this would require some knowledge about the distribution of $X_1, \ldots, X_n$. To make things rigorous, we assume that $X_1, \ldots, X_n$ are the *binary* (*fractional*) *representations of independent random variables with common density* f *on* $[0, 1)$. In

0010-485X/82/0028/0367/$01.00

that case, A_n too is a random variable, with expected value

$$E(A_n)=\frac{1}{n}\sum_{i=1}^{n} E(D_{ni})=E(D_{n1}) \quad \text{(by symmetry)}.$$

A trie with n leaves has at least $2n-1$ nodes, and the average distance to the root (A_n) is at least equal to the average distance to the root of the leaves in a complete binary tree with $2n-1$ nodes, and thus,

$$A_n \geq [\log_2(2n-1)]-1 \sim \log_2 n. \tag{1}$$

The actual value of A_n increases as the X_i's become more clustered. Smooth distributions of the X_i's lead to lower values of A_n. In this note, we will see to what extent the distribution of the X_i's influences $E(A_n)$.

Theorem 1: *If f is the uniform distribution on $[0,1)$, then*

$$0 \leq E(A_n)-\log_2(n-1)-\frac{\gamma}{\log 2} \leq 1+\frac{1}{(2n-2)\log 2}, \quad n \geq 2,$$

where $\gamma=0.5572156649\ldots$ is Euler's constant. Thus,

$$\lim_n E(A_n)/\log_2 n=1.$$

Thus, the expected average depth of a trie varies as $\log_2 n$ for uniform distributions, and in view of (1), this is the optimal asymptotic rate. The same result can also be found in Knuth (1973b), but we include a new short proof anyway. The main result of this note is Theorem 2, where Theorem 1 is generalized towards *all* densities on $[0,1)$.

Theorem 2: *Let f be a density on $[0,1)$. Then either*

$$E(A_n)=\infty \quad \textit{for all } n \geq 2,$$

or

$$\lim_n E(A_n)/\log_2 n=1$$

according to whether $\int f^2(x)\,dx=\infty$ or $\int f^2(x)\,dx<\infty$.

Theorem 2 states that either tries are on the average asymptotically optimal $\left(\lim_n E(A_n)/\log_2 n=1\right)$ or they are bad for all n $\left(\inf_{n\geq 2} E(A_n)=\infty\right)$. There are no intermediate situations. The crucial condition is the square integrability of f (which is a condition on the peak(s) of the density). Theorem 2 offers at the same time a nice characterization of densities that are square integrable: $\int f^2(x)\,dx<\infty$ if and only if $E(A_2)<\infty$ (i.e. if and only if the expected length of the largest common left substring of X_1 and X_2 is finite).

We remark that the second statement of Theorem 2 follows from (1) and the inequality

$$E(A_n) \leq \log_2 n+1+\left(\gamma+\frac{1}{2n-2}\right)/\log 2+192\int f^2(x)\,dx, \tag{2}$$

valid for all $n \geq 2$. Inequality (2) is not tight, but suffices to prove the Theorem. Also, notice that $\int f^2(x)\,dx$ influences only the constant term, and not the coefficient of $\log_2 n$.

We notice finally that no continuity conditions are imposed on f in Theorem 2. This will force us to use some advanced measure theoretical tools in the proof.

2. Proofs

Partition $[0, 1)$ into sets

$$A_{ki} = \left\{x : \frac{i-1}{2^k} \leq x < \frac{i}{2^k}\right\}, \quad 1 \leq i \leq 2^k.$$

If $x, y \in A_{ki}$, then the first k bits in the binary fractions of x and y are identical. We let $A_k(x)$ be the set A_{ki} to which x belongs, and define the function g_k by

$$g_k(x) = \int_{A_k(x)} f(y)\,dy.$$

Then,

$$E(A_n) = E(D_{n1}) = \sum_{k=0}^{\infty} P(D_{n1} > k)$$

$$= \sum_{k=0}^{\infty} \int f(x)\; P\left(\bigcup_{j=2}^{n} [\text{first } k \text{ bits of } X_j \text{ and } x \text{ are identical}]\right) dx \tag{3}$$

$$= \sum_{k=0}^{\infty} \int f(x)\,(1-(1-g_k(x))^{n-1})\,dx.$$

Formula (3) will be our starting point.

Proof of Theorem 1:
(3) is equal to

$$\sum_{k=0}^{\infty} (1-(1-2^{-k})^{n-1})$$

for the uniform distribution. This quantity in turn lies between a_n and a_n+1 where

$$a_n = \int_0^{\infty} (1-(1-2^{-x})^{n-1})\,dx.$$

By the transformation

$$1-2^{-x} = y\,(2^x = (1-y)^{-1};\; x = -\log_2(1-y);\; dx = dy/(1-y)\log 2)$$

we see that

$$a_n = \int_0^1 (\log 2)^{-1}\,\frac{1-y^{n-1}}{1-y}\,dy = \int_0^1 (\log 2)^{-1}(1+y+\ldots+y^{n-2})\,dy$$

$$= (\log 2)^{-1} \sum_{i=1}^{n-1} \frac{1}{i}.$$

By using inequalities for the harmonic series (Knuth, 1973a, pp. 74, 111) we see that $a_n \log 2$ lies between $\gamma+\log(n-1)$ and $\gamma+\log(n-1)+\frac{1}{2n-2}$, $n\geq 2$. This completes the proof of Theorem 1.

Proof of Theorem 2:
For $n\geq 2$, we have from (3)

$$E(A_n)\geq \sum_{k=0}^{\infty} \int f(x)(1-1+g_k(x))\,dx=\sum_{k=0}^{\infty} \int f(x)\,g_k(x)\,dx.$$

But by Fatou's lemma,

$$\begin{aligned}\liminf_k 2^k \int f(x)\,g_k(x)\,dx &\geq \int f(x) \liminf_k 2^k g_k(x)\,dx\\ &=\int f^2(x)\,dx\end{aligned}$$

where we also used the fact that for almost all x, $\lim_k 2^k g_k(x)=f(x)$ (Lebesgue density theorem (see Wheeden and Zygmund, 1977); also derivable from the martingale convergence theorem (see Breiman, 1968)). Thus,

$$\inf_{n\geq 2} E(A_n)\geq \sum_{k=0}^{\infty} 2^{-k}\left(\int f^2(x)\,dx+o(1)\right)=\infty$$

when $\int f^2(x)\,dx=\infty$. Theorem 2 now follows if we can show (2).

We introduce the Hardy-Littlewood maximal function (see Wheeden and Zygmund, 1977, pp. 155)

$$f^*(x)=\sup_{r>0}(2r)^{-1} \int_{|y-x|<r} f(y)\,dy.$$

It is clear that

$$\begin{aligned}2^k g_k(x) &\leq \sup_{r>0} \max\left(\frac{1}{r}\int_{0<y-x<r} f(y)\,dy,\ \frac{1}{r}\int_{-r>y-x<0} f(y)\,dy\right)\\ &\leq \sup_{r>0} 2\left((2r)^{-1}\int_{|y-x|<r} f(y)\,dy\right)=2f^*(x).\end{aligned}$$

From (3),

$$\begin{aligned}E(A_n) \leq &\sum_{k=0}^{\infty} \int_{f^*(x)<2^{k-1}} f(x)\left(1-(1-f^*(x)/2^{k-1})^{n-1}\right)dx\\ &+\sum_{k=0}^{\infty} \int_{f^*(x)\geq 2^{k-1}} f(x)\,dx.\end{aligned} \tag{4}$$

The last term in (4) does not exceed

$$\begin{aligned}\sum_{k=0}^{\infty} \int f(x)\,f^*(x)/2^{k-1}\,dx &\leq \sum_{k=0}^{\infty} \int f^{*2}(x)/2^{k-1}\,dx\\ &=4\int f^{*2}(x)\,dx\leq 192\int f^2(x)\,dx\end{aligned} \tag{5}$$

where we first used Chebyshev's inequality, and then used an inequality between the integrals of f^{*2} and f^2 (see Wheeden and Zygmund, 1977, pp. 156, and derive the constants by carefully analyzing Vitali's lemma (pp. 102) and the Hardy-Littlewood inequality (pp. 105)).

Consider now the first term on the right-hand-side of (4), and note that

$$\sum_{\substack{k=0 \\ 2^{k-1} > f^*(x)}}^{\infty} \left(1-\left(1-f^*(x)/2^{k-1}\right)^{n-1}\right)$$
$$\leq 1+ \int_{2^y > 2f^*(x)} \left(1-\left(1-2f^*(x)/2^y\right)^{n-1}\right) dy$$
$$= 1+a_n$$

where a_n is defined as in the proof of Theorem 1 (the last step follows from the transformation $z=1-2f^*(x)/2^y$, $dy=dz/(1-z)\log 2$). Since $1+a_n$ does not depend upon x, we see that (4) is bounded from above by $1+a_n+192\int f^2(x)\,dx$. Inequality (2) follows from the inequalities for a_n derived in the proof of Theorem 1.

Acknowledgement

The present problem was brought to the author's attention through discussions with Tim Merrett and Jack Orenstein at McGill University.

References

[1] Breiman, L.: Probability. Reading, Mass.: Addison-Wesley 1968.
[2] Fredkin, E. H.: Trie memory. Communications of the ACM *3*, 490—500 (1960).
[3] Knuth, D. E.: The art of computer programming; Vol. 1: Fundamental algorithms. Reading, Mass.: Addison-Wesley 1973a.
[4] Knuth, D. E.: The art of computer programming; Vol. 3: Searching and sorting. Reading, Mass.: Addison-Wesley 1973b.
[5] Wheeden, R. L., Zygmund, A.: Measure and integral. New York: Marcel Dekker 1977.

L. Devroye
School of Computer Science
McGill University
805 Sherbrooke Street West
Montreal
Canada H3A 2K6