International Journal of Pure and Applied Mathematics
Volume 80 No. 1 2012, 53-61
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu
ijpam.eu

ON AVERAGE PROFILE OF THE BINARY BUCKET DIGITAL SEARCH TREES

Ramin Kazemi

Department of Statistics
Imam Khomeini International University
Qazvin, IRAN

Abstract: Drmota and Szpankowski have studied the average profile in digital search trees (DST) [2]. In this paper, we extend the same approach to bucket digital search trees (b-DSTs) where each node can hold up to b keys. The construction rule of b-DSTs is the same as DSTs, except that keys keep staying in a node as long as its capacity remains less than b. Here we apply an alternate but unified and shorter approach to the analysis of the expectation of two random variables (internal and external profiles) in b-DSTs. We show that the asymptotic results are independent of b and are quite equal to the average profile of ordinary digital search trees in spite of the partial differential equations arising here are of the order b.

AMS Subject Classification: 05C05
Key Words: bucket digital search trees, external and internal profiles, average profiles

1. Introduction

Digital search trees for $b = 1$ have been analyzed in the past in the case of a fixed number of independent strings (see the references in [6]). Much less is known

Received: May 8, 2012

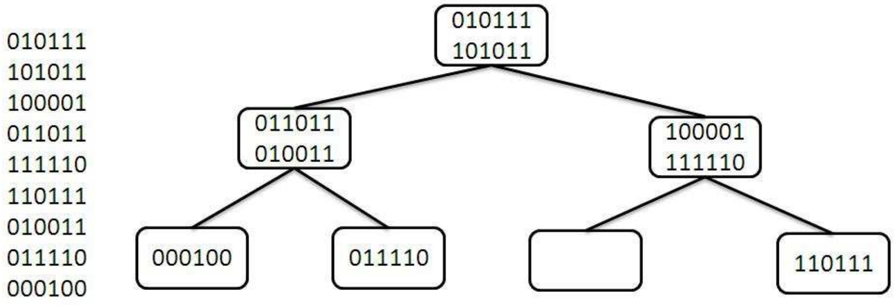

Figure 1: A 2-DST with 9 keys

about b-digital search trees. In binary b-digital search trees or generalized digital search trees [5] with bucket capacity b, groups of up to b keys are stored in the nodes, and branching to the subtrees is based on a digital search method. These trees are used to store keys taken from some alphabet. We consider in this paper the binary alphabet. Thus the keys considered will be viewed as sequences of binary digits from the set $\Sigma = \{0, 1\}$. In binary b-digital search trees, the first b keys are stored in the root node; a subsequent key is guided to the left subtree if its first symbol is 0, or to the right if the first symbol is 1. The subtrees are recursively constructed by the same algorithm, but if the subtree root is at level i, the $(i + 1)$st bit of the key is used for branching. See Figure 1 for an illustration.

The probability model usually assumed for the analysis of digital trees is the Bernoulli model (each string is a binary i.i.d. sequence with p being the probability of a "1" $(0 < p < 1)$). This simple model may seem too idealized for practical purposes (see the references in [2]). By assuming that the n input strings are independent and follow a binary Bernoulli model the asymptotic behaviour of the average profile in ordinary digital search trees was determined by Drmota and Szpankowski [2].

It was already observed by Flajolet and Richmond [3] that b-digital search trees are harder to analyze than ordinary digital search trees. The difficulty boils dowan ultimately to a solution of the general recurrence. Louchard, Szpankowski and Tang have analyzed b-digital search trees to derive the expected number of strings on level k [6]. They also investigated the depth of a randomly selected node in such a tree. Hubalek, Hwang, Lew, Mahmoud and Prodinger [5] toke a multivariate view of digital search trees by studying the number of nodes of different types that may coexist in a b-digital search tree as it grows under an arbitrary memory management system. They obtained the mean of each type of node, as well as the entire covariance matrix between types, whereupon weak laws of large numbers follow from the orders of magnitude. Furthermore, They used a method of moments to show that the distribution is asymptotically normal. The method of proof there is of some generality and is applicable to other parameters like path length and size in random tries and Patricia tries.

We always call a node v with capacity $c = b$, *complete* and otherwise *incomplete*. Let I_n^k be the random number of complete nodes at level k in a binary b-digital search tree built over n strings generated by a Bernoulli model with parameter $p < q$. Also let B_n^k be the random number of incomplete nodes and available nodes which are directly attached to complete nodes already existing in the tree at level k. These definitions are a generalization of *internal* and *external* profiles in ordinary digital search trees. Here we also call these quantities profiles. Let $q := 1 - p$ and $p < q$. We also mention that symmetric b-DST's, i.e. $p = \frac{1}{2}$, are not covered by our analysis because the saddle point analysis fails for $p = q$ [2].

2. Exponential Generating Functions

In this section we first derive a general formula for the generating functions of the external and internal profiles. Then we discuss on the asymmetric b-digital search trees; i.e., $p < q$.

Suppose that there are $n+b$ strings to store. The root of such a tree contains b strings and the remaining n strings are split between the left subtree and the right subtree. If ℓ strings go to the left subtree, then its probability generating function is characterized by $\phi_{B_{l,k}}(u) = \mathbf{E}\,[u^{B_\ell^k}]$ while $\phi_{B_{n-l,k}}(u) = \mathbf{E}\,[u^{B_{n-\ell}^k}]$ is the probability generating function for the right subtree. Finally, the probability generating function of the external profile, satisfies the following recurrence relation

$$\phi_{B_{n+b,k+1}}(u) = \sum_{\ell=0}^{n} \binom{n}{\ell} p^\ell q^{n-\ell} \phi_{B_{l,k}}(u)\phi_{B_{n-l,k}}(u). \tag{1}$$

The functional recurrence (1) translates into

$$\frac{d^b}{dx^b} W_k(x,u) = W_{k-1}(px,u)W_{k-1}(qx,u), \qquad (k \geq 1), \tag{2}$$

where

$$W_k(x,u)=\sum_{n\geq 0}\phi_{B_{n,k}}(u)\frac{x^n}{n!}$$

and $W_0(x,u)=u+e^x-1$ and $W_k(0,u)=1$ $(k\geq 1)$. Similarly, the corresponding generating function for the internal profile

$$\overline{W}_k(x,u)=\sum_{n\geq 0}\phi_{I_{n,k}}(u)\frac{x^n}{n!}$$

satisfies the same recurrence relation

$$\frac{d^b}{dx^b}\overline{W}_k(x,u)=\overline{W}_{k-1}(px,u)\overline{W}_{k-1}(qx,u),\qquad (k\geq 1),\tag{3}$$

with the initial conditions $\overline{W}_0(x,u)=1+u(e^x-1)$ and $\overline{W}_k(0,u)=1$ $(k\geq 1)$.

We are interested in the expected profiles $\mathbf{E}\,[B_n^k]$ and $\mathbf{E}\,[I_n^k]$. By taking derivatives with respect to u and setting $u=1$ we obtain for the exponential generating function

$$E_k(x)=\sum_{n\geq 0}\mathbf{E}\,[B_n^k]\,\frac{x^n}{n!}$$

the following functional recurrence

$$\frac{d^b}{dx^b}E_k(x)=e^{qx}E_{k-1}(px)+e^{px}E_{k-1}(qx),\tag{4}$$

with initial condition $E_0(x)=1$ and $E_k(0)=0$ $(k\geq 1)$. The corresponding generating function for the internal profile

$$\overline{E}_k(x)=\sum_{n\geq 0}\mathbf{E}\,[I_n^k]\,\frac{x^n}{n!}$$

satisfies recurrence (4), too, however with initial conditions $\overline{E}_0(x)=e^x-1$ and $\overline{E}_k(0)=0$ $(k\geq 1)$.

The Poisson transform of $E_k(x)$, namely $\Delta_k(x)=e^{-x}E_k(x)$ translates recurrence (4) into

$$e^{-x}\frac{d^b}{dx^b}E_k(x)=\Delta_{k-1}(px)+\Delta_{k-1}(qx),\qquad (k\geq 1).\tag{5}$$

It is obvious that [2]

$$\Delta_k(x)+\Delta_k'(x)=\Delta_{k-1}(px)+\Delta_{k-1}(qx).$$

It is easy to show that

$$e^{-x}\frac{d^b}{dx^b}E_k(x)=\sum_{j=0}^{b}\binom{b}{j}\frac{d^j}{dx^j}\Delta_k(x).$$

Thus

$$\Delta_k(x)+\sum_{j=1}^{b}\binom{b}{j}\frac{d^j}{dx^j}\Delta_k(x)=\Delta_{k-1}(px)+\Delta_{k-1}(qx) \tag{6}$$

with initial conditions $\Delta_0(x)=e^{-x}$ and $\Delta_k(0)=0$ $(k\geq 1)$. It is easy to prove (by induction) that $\Delta_k(x)$ can be represented as a finite linear combination of functions of the form of

$$\exp\Big\{-\beta_1(p)^{\ell_1}\beta_2(q)^{-\ell_2}x\Big\}$$

with $\beta_1(p),\beta_2(q)$ are functions of p and q, respectively and $0\leq \ell_1+\ell_2\leq k$. For internal profile, $\overline{\Delta}_0(x)=1-e^{-x}$ and $\Delta_k(0)=0$ $(k\geq 1)$. Let $\overline{\Delta}_k^*(s)$ be the Mellin transform of $\Delta_k(x)$ and $T(s)=p^{-s}+q^{-s}$. The recurrence (6) can be translated into

$$\Delta_k^*(s)+\sum_{j=1}^{b}\binom{b}{j}(-1)^j(s-1)^j\Delta_k^*(s-j)=T(s)\Delta_{k-1}^*(s). \tag{7}$$

We can express $\Delta_k^*(s)$ as [4]

$$\Delta_k^*(s)=\Gamma(s)F_k(s),$$

where $\Gamma(s)$ is the Euler gamma function. In the above, $F_k(s)$ is the finite linear combinations of functions of

$$\beta_1(p)^{-\ell_1 s}\beta_2(q)^{-\ell_2 s}.$$

It is clear that (7) translates into

$$\begin{aligned} F_k(s) \quad &+ \quad \sum_{j=1}^{b}\binom{b}{j}(-1)^j\frac{(s-1)^j}{(s-1)\cdots(s-j)}F_k(s-j) \\ &= \quad T(s)F_{k-1}(s) \end{aligned} \tag{8}$$

with initial condition $F_0(s)=1$. These recurrence is satisfied for internal profile, too.

In order to find a solution of (8) we define the power series $f(s,w)=\sum_{k\ge 0}F_k(s)w^k$ that translates (8) into

$$f(s,w)=\frac{\sum_{j=1}^{b}\binom{b}{j}(-1)^{j+1}\frac{(s-1)^j}{(s-1)\cdots(s-j)}f(s-j,w)}{1-wT(s)}. \tag{9}$$

3. The Main Results

If we define the partial functional operator $\mathbf{A}$ as

$$\mathbf{A}[h](s)=\sum_{j\ge 0}h(s-j)T(s-j)$$

for some function h, then

$$F_k(s)=\mathbf{A}[F_{k-1}](s)-\mathbf{A}[F_{k-1}](0) \qquad (k\ge 1) \tag{10}$$

just similar to [2]. In this section we apply an alternate but unified (and shorter) approach to the analysis of the expectation of profiles in b-DSTs. We show that the asymptotic results are independent of b and are quite equal to the average profile of ordinary digital search trees in spite of the partial differential equations arising here are of the order b. We also set $R_k(s)=\mathbf{A}^k[1](s)$. These functions have a general representation and satisfies the recurrence

$$\begin{aligned} R_{k+1}(s)-R_{k+1}(s-1) &= \mathbf{A}[R_k](s)-\mathbf{A}[R_k](s-1) \\ &= T(s)R_k(s). \end{aligned} \tag{11}$$

Also $R_k(-\infty)=0$ $(k\ge 1)$ [2].

Lemma 1. *Suppose $g(s,w)=\sum_{k\ge 0}R_k(s)w^k$, then*

$$g(s,w)=\frac{1}{\prod_{j\ge 0}(1-wT(s-j))} \tag{12}$$

Proof. We have

$$\begin{aligned} g(s,w)-g(s-1,w) &= \sum_{k\ge 0}\Big(R_k(s)-R_k(s-1)\Big)w^k \\ &= wT(s)\sum_{k\ge 0}R_k(s)w^k \end{aligned}$$

$$= \quad wT(s)g(s,w).$$

or $g(s,w) = g(s-1,w)/(1-wT(s))$ and consequently

$$g(s,w) = \frac{1}{\prod_{j\geq 0}(1-wT(s-j))}$$

because $g(-\infty,w) = 1$. □

Lemma 1 shows that $w = 1/T(s)$ is the dominanting polar singularity of $g(s,w)$ if s is sufficiently close to the real axis. Also

$$F_k(s) = R_k(s) - \sum_{\ell=0}^{k-1} F_\ell(s)R_{k-\ell}(a), \qquad (k \geq 0). \tag{13}$$

where for external profile $a = 0$ and for internal profile $a = -1$ (of course, for internal nodes: $\overline{\Delta}_k^*(s) = -\Gamma(s)F_k(s)$).

We will prove (13) by induction. Certainly, it is satisfied for $k = 0$. Now suppose that is holds for some $k \geq 0$. By (10) we also have

$$\begin{aligned} F_{k+1}(s) &= \mathbf{A}[F_k](s) - \mathbf{A}[F_k](a) \\ &= \mathbf{A}[R_k](s) - \mathbf{A}[R_k](a) \\ &- \sum_{\ell=0}^{k-1}(\mathbf{A}[F_\ell](s) - \mathbf{A}[F_\ell](a))R_{k-\ell}(a) \\ &= R_{k+1}(s) - R_{k+1}(a) \\ &- \sum_{\ell=0}^{k-1} F_{\ell+1}(s)R_{k-\ell}(a) \\ &= R_{k+1}(s) - \sum_{\ell=0}^{k} F_\ell(s)R_{k+1-\ell}(a). \end{aligned}$$

Finally, since $F_k(s) = \pm\Delta_k^*(s)/\Gamma(s)$ is analytic and $1/\Gamma(-\ell) = 0$, it also follows that $F_k(-\ell) = 0$ for $\ell = 0, 1, \ldots, k-1$.

Theorem 1. *The power series $f(s,w)$ is given by*

$$f(s,w) = \frac{g(s,w)}{g(a,w)} \tag{14}$$

and as $k \to \infty$,

$$F_k(s) \quad \sim \quad \frac{\prod_{j\geq 0}(1 - T(-j)/T(s))}{\prod_{\ell\geq 1}(1 - T(s-\ell)/T(s))}\, T(s)^k.$$

Proof. The formal identity (14) is hold because

$$f(s,w)g(a,w)=\sum_{k\geq 0}\ \sum_{\ell=1}^{k}F_\ell(s)R_{k-\ell}(a)w^k,$$

Note that the mapping $w\mapsto 1/g(a,w)$ is an entire function. Hence, $w=1/T(s)$ is again the dominating polar singularity if s is sufficiently close to the real axis. Consequently

$$\begin{aligned}
F_k(s) &= [w^k]f(s,w)\\
&\sim -Res\left[\frac{g(s,w)}{g(0,w)}w^{-k-1};w=\frac{1}{T(s)}\right]\\
&= -\lim_{w\to\frac{1}{T(s)}}\left(w-\frac{1}{T(s)}\right)\frac{\prod_{\ell\geq 0}\frac{1}{1-wT(s-\ell)}}{\prod_{j\geq 0}\frac{1}{1-wT(-j)}}w^{-k-1}\\
&= \lim_{w\to\frac{1}{T(s)}}\frac{1-wT(s)}{T(s)}\frac{\prod_{\ell\geq 0}\frac{1}{1-wT(s-\ell)}}{\prod_{j\geq 0}\frac{1}{1-wT(-j)}}w^{-k-1}\\
&= \frac{\prod_{j\geq 0}(1-T(-j)/T(s))}{\prod_{\ell\geq 1}(1-T(s-\ell)/T(s))}T(s)^k.
\end{aligned}$$

□

4. Conclusion

From Theorem 1, we know that $\Delta_k^*(s)=\pm\Gamma(s)F_k(s)$ behave asymptotically as $T(s)^k$. Thus we are in the same situation as in the analysis of the average profile presented in [2] for ordinary digital search trees. However, We review a very short outline of the proof. In order to estimate complex integrals of $\Delta_k(n)$, we adopt the contour of integration a curve that crosses the saddle points ρ of the integrand [2]. Since integral is dependant on the large parameter n, this strategy provides a accurate asymptotic information. The final step in the proof is to obtain asymptotics for $\mathbf{E}\,[B_n^k]$ and $\mathbf{E}\,[I_n^k]$ from the asymptotic properties of $\Delta_k(x)$ and $\overline{\Delta}_k(x)$. This is accomplished by the analytical depoissonization [4] which requires to compute Cauchy integrals

$$\mathbf{E}\,[B_n^k] \quad = \quad \frac{n!}{2\pi i}\int_{|x|=n}e^x\Delta_k(x)\frac{dx}{x^{n+1}},$$

$$\mathbf{E}\,[I_n^k] \;=\; \frac{n!}{2\pi i}\int_{|x|=n} e^x\overline{\Delta}_k(x)\frac{dx}{x^{n+1}}.$$

Since $\Delta_k(x)$ and $\overline{\Delta}_k(x)$ behave quite smoothly (in particular they have a subexponential growth) the depoissonization heuristics saying that $\mathbf{E}\,[B_n^k] \approx \Delta_k(n)$ and $\mathbf{E}\,[I_n^k] \approx \overline{\Delta}_k(n)$ apply. However, there is a precise analysis in [2].

References

[1] M. Drmota, *Random Trees, An Interplay Between Combinatorics and Probability*, Springer, Wien-New York (2009).

[2] M. Drmota, W. Szpankowski, The expected profile of digital search trees, *Journal of Combinatorial Theory, Series A*, **118** (2011), 1939-1965.

[3] P. Flajolet, B. Richmond, Generalized digital trees and their difference-differential equations, *Random Structure and Algorithms*, **3**, No. 3 (1992), 305-320.

[4] P. Flajolet, R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge (2008).

[5] F. Hubalek, H.K. Hwang, W. Lew, H. Mahmoud, H. Prodinger, A multivariate view of random bucket digital search trees, *Journal of Algorithms*, **44**, No. 1 (2002), 121-158.

[6] G. Louchard, W. Szpankowski, J. Tang, Average profile of the generalized digital search tree and the generalized Lempel-Ziv algorithm, *SIAM J. Comput.*, **28**, No. 3 (1999), 904-934.