International Journal of Pure and Applied Mathematics
Volume 80 No. 5 2012, 665-672
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

(b, m)-DIGITAL SEARCH TREES

Ramin Kazemi

Department of Statistics
Imam Khomeini International University
Qazvin, IRAN

Abstract: In this paper we study the m-ary digital search trees ($m \geq 2$) with maximal bucket size b (≥ 1). This model can be considered as a simultaneous extension of ordinary digital search trees (or $(1, 2)$-digital search trees in our notation). We construct this model by using strings over an alphabet leading to m-ary trees and staying strings in a node as long as its capacity remains less than b. We obtain the exact formulas for the mean of the profiles of nodes in symmetric case. Also we discuss on asymptotic of mean in asymmetric case. Our analysis for $b = 1$ and $m = 2$ reduce to the previous analysis on ordinary digital search trees.

AMS Subject Classification: 05C05
Key Words: (b, m)-digital search trees, profiles

1. Introduction

Digital search trees for $b = 1$ and binary alphabet ($m = 2$) have been analyzed in the past (see the references in [8]). We consider in this paper the m-ary alphabet ($m \geq 2$). Thus the keys considered will be viewed as sequences of m-ary digits from the set $\mathcal{A} = \{0, 1, 2, ..., m\}$. In (b, m)-digital search trees, the first b keys are stored in the root node; a subsequent key is guided to the each of m subtrees according to its first symbol from left to right. Also keys keep staying in a node as long as its capacity remains less than b. The subtrees are recursively constructed by the same algorithm, but if the subtree root is at

Received: July 8, 2012

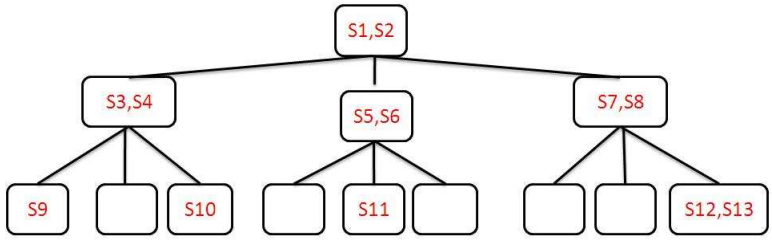

Figure 1: A (2,3a)-digital search tree with 13 strings (S_1 = 01020..., S_2 = 00129..., S_3 = 01020..., S_4 = 01101..., S_5 = 10112..., S_6 = 12011..., S_7 = 20010..., S_8 = 21001..., S_9 = 00020..., S_{10} = 02100..., S_{11} = 11000..., S_{12} = 22010..., S_{13} = 22101...).

level ℓ, the $(\ell+1)$st bit of the key is used for branching. See Figure 1 for an illustration of $(2,3)$-digital search tree with 13 strings.

In this paper we study the (b,m)-digital search trees built over n m-ary strings generated by an extended memoryless source (see the references in [1]). More precisely, we assume each string is a m-ary i.i.d. sequence with p_i being the probability of a"i" $(i = 1, ..., m)$. The corresponding (b,m)-digital search trees constructed from these n strings is called a random (b,m)-digital search trees. For the symmetric case, $p_i = 1/m$ $(i = 1, ..., m)$.

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 (b,m)-digital search tree built over n strings generated by a memoryless source with parameters $0 < p_m < p_{m-1} < \cdots < p_1 < 1$. 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 $(1,2)$-digital search trees.

The asymptotic behaviour of the average profile in symmetric and asymmetric $(1,2)$- digital search trees were determined by Drmota and Szpankowski [1]. Louchard obtained an exact and asymptotic distributions in $(1,2)$-digital Search Trees [5]. Louchard and et.al., have analyzed $(b,2)$-digital search trees to derive the expected number of strings on level k [8]. They also investigated the depth of a randomly selected node in such a tree. Hubalek and et. al., toke a multivariate view of digital search trees by studying the number of nodes of different types that may coexist in a $(b,2)$-digital search tree as it grows under an arbitrary memory management system [3]. It should be noted that the explicit formula for $\mathbf{E}[B_n^k]$ has appeared several times in [5, 6, 7] for $(1,2)$-digital

search trees. In this paper we extend these results to $\mathbf{E}\,[I_n^k]$ in (b,m)-digital search trees that cover the previous results.

2. Preliminaries

In this section we derive a general formula for the generating functions of the profiles. 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 subtrees. If k_j strings go to the each subtree, then its probability generating function is characterized by $\phi_{k_j,k}(u) = \mathbf{E}\,[u^{B_{k_j}^k}]$. Finally, the probability generating function of the external profile, satisfies the following recurrence relation

$$\begin{aligned}\phi_{n+b,k+1}(u) &= \sum_{k_i\geq 0}^{n} \left(\begin{array}{c} n \\ k_1,...,k_{m-1} \end{array} \right) p_1^{k_1}\cdots p_{m-1}^{k_{m-1}} p_m^{n-\sum_{i=1}^{m-1} k_i} \\ &\cdot \quad \phi_{k_1,k-1}(u)\cdots\phi_{k_{m-1},k-1}(u)\phi_{n-\sum_{i=1}^{m-1} k_i,k-1}(u). \qquad (1)\end{aligned}$$

Due to the independence assumption the number of strings where the first digits are $1,...,m-1$ follows a multinomial distribution

$$M(n,p_1,...,p_{m-1}).$$

The splitting probabilities are thus given by

$$\left(\begin{array}{c} n \\ k_1,...,k_{m-1} \end{array} \right) p_1^{k_1}\cdots p_{m-1}^{k_{m-1}} p_m^{n-\sum_{i=1}^{m-1} k_i}.$$

The functional recurrence (1) translates into

$$\frac{d^b}{dx^b} W_k(x,u) = \prod_{i=1}^{m} W_{k-1}(p_i x,u), \qquad (k\geq 1), \qquad (2)$$

where

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

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

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

satisfies the same recurrence relation

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

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

By taking derivatives with respect to u and setting $u=1$ we obtain for the exponential generating function

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

the following functional recurrence

$$\frac{d^b}{dx^b}E_k(x)=\sum_{i=1}^{m}e^{p_ix}E_{k-1}(p_ix) \tag{4}$$

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

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

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

3. Symmetric Case

Let us start with the symmetric case $p_i=\frac{1}{m},\ (i=1,2,...,m)$. The corresponding generating functions have simpler structures. Namely

$$\begin{aligned}\frac{d^b}{dx^b}W_k(x,u) &= W_{k-1}\Big(\frac{x}{m},u\Big)^m,\\ \frac{d^b}{dx^b}\overline{W}_k(x,u) &= \overline{W}_{k-1}\Big(\frac{x}{m},u\Big)^m,\qquad (k\geq 1)\end{aligned}$$

and

$$\frac{d^b}{dx^b}E_k(x) = me^{x/m}E_{k-1}\Big(\frac{x}{m}\Big), \tag{5}$$

$$\frac{d^b}{dx^b}\overline{E}_k(x) \quad = \quad me^{x/m}\overline{E}_{k-1}\Big(\frac{x}{m}\Big). \tag{6}$$

Set $Q_{0,m,b}=1$ and

$$Q_{\ell,m,b}=\prod_{j=1}^{\ell}\left(1-m^{-j}\right)^b, \qquad (\ell>0, b\geq 1, m\geq 2).$$

Lemma 1. *For $k\geq 0$,*

$$E_k(x)=m^k e^x\sum_{j=0}^{k}\frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}e^{-xm^{j-k}} \tag{7}$$

and

$$\overline{E}_k(x)=m^k e^x\left(1-\sum_{j=0}^{k}\frac{(-1)^j m^{-\binom{j+1}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}e^{-xm^{j-k}}\right). \tag{8}$$

Proof. We will prove these relations by induction. Certainly, they are satisfied for $k=0$ (and $b=1, m=2$) [5, 6, 7]. Now suppose that they hold for some $k\geq 0$. Thus from (5),

$$\begin{aligned} E_{k+1}(x) &= m^{k+1}\sum_{j=0}^{k+1}\left(1-m^{j-(k+1)}\right)^{-b}\frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}e^{(1-m^{j-(k+1)})x} \\ &= m^{k+1}e^x\sum_{j=0}^{k+1}\frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k+1-j,m,b}}e^{-xm^{j-(k+1)}}. \end{aligned}$$

The equality (8) is established in the same manner. □

Theorem 1. *For any n and $k\leq n$,*

$$\mathbf{E}[B_n^k]=m^k\sum_{j=0}^{k}\frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}\left(1-\frac{1}{m^{k-j}}\right)^n \tag{9}$$

and

$$\mathbf{E}[I_n^k]=m^k\left[1-\sum_{j=0}^{k}\frac{(-1)^j m^{-\binom{j+1}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}\left(1-\frac{1}{m^{k-j}}\right)^n\right]. \tag{10}$$

Proof. Let $[x^n]f(x)$ denote the operation of extracting the coefficient of x^n in the formal power series $f(x) = \sum f_n x^n$. We have $[x^n]f(qx) = q^n[x^n]f(x)$. Thus

$$\begin{aligned}\mathbf{E}[B_n^k] &= n![x^n]E_k(x)\\ &= n!m^k \sum_{j=0}^{k} \frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k-j,m,b}}[x^n]e^{(1-m^{j-k})x}\\ &= m^k \sum_{j=0}^{k} \frac{(-1)^j m^{-\binom{j}{m}}}{Q_{j,m,b}Q_{k-j,m,b}} \left(1-\frac{1}{m^{k-j}}\right)^n.\end{aligned}$$

The equality (10) is established in the same manner. □

We see that $\mathbf{E}[B_n^k]$ and $\mathbf{E}[I_n^k]$ are dependent on b and m in symmetric case.

4. Asymmetric Case

The Poisson transform of $E_k(x)$, namely $\Delta_k(x) = e^{-x}E_k(x)$ translates recurrence (4) into [2]

$$e^{-x}\frac{d^b}{dx^b}E_k(x) = \sum_{i=1}^{m} \Delta_{k-1}(p_i x), \qquad (k \geq 1). \tag{11}$$

It is easy to prove 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

$$\Delta_k(x) + \sum_{j=1}^{b} \binom{b}{j} \frac{d^j}{dx^j}\Delta_k(x) = \sum_{i=1}^{m} \Delta_{k-1}(p_i x) \tag{12}$$

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 combinations of functions of the form of $\exp\{-\beta_i(p_i)^{\ell_i}x\}$ with $\beta_i(p_i)$ are functions of p_i and $0 \leq \sum_{i=1}^{k} \ell_i \leq k$. 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 $T_m(s) = p_1^{-s} + p_2^{-s} + \cdots + p_m^{-s}$. The recurrence (12) can be translated into

$$\Delta_k^*(s) + \sum_{j=1}^{b} \binom{b}{j} (-1)^j (s-1)^j \Delta_k^*(s-j) = T_m(s)\Delta_{k-1}^*(s). \tag{13}$$

We can express $\Delta_k^*(s)$ as $\Delta_k^*(s) = \Gamma(s)F_k(s)$, where $\Gamma(s)$ is the Euler gamma function. In the above, $F_k(s)$ is the finite linear combinations of functions of $\prod_{i=1}^m \beta_i(p_i)^{-\ell_i s}$. It is obvious that (13) translates into

$$\begin{aligned} F_k(s) &+ \sum_{j=1}^{b} \binom{b}{j}(-1)^j \frac{(s-1)^j}{(s-1)\cdots(s-j)} F_k(s-j) \\ &= T_m(s)F_{k-1}(s) \end{aligned} \tag{14}$$

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

In order to find a solution of (14) we define the power series $f(s,w) = \sum_{k\geq 0} F_k(s)w^k$ that translates (14) into

$$f(s,w) = \frac{\sum_{j=1}^{b} \binom{b}{j} \frac{(-1)^{j+1}(s-1)^j}{(s-1)\cdots(s-j)} f(s-j,w)}{1 - wT_m(s)}. \tag{15}$$

If we define $\mathbf{A}[h](s) = \sum_{j\geq 0} h(s-j)T(s-j)$ for some function h, then we can show that the asymptotic results are independent of b and m and are quite equal to the average profile of $(1,2)$-digital search trees in spite of the partial differential equations arising here are of the order b just similar to [4].

References

[1] M. Drmota, W. Szpankowski, The expected profile of digital search trees, *J. Combin. Theory Ser. A.*, **118** (2011), 1939-1965.

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

[3] 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.

[4] R. Kazemi, On average profile of the binary bucket digital search trees, *Int. J. Pure and Appl. Math.*, **80**, No. 1 (2012), In Press.

[5] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *RAIRO Theoretical Inform. Applications*, **21** (1987), 479-495.

[6] G. Louchard, W. Szpankowski, Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm, *IEEE Trans. Information Theory*, **41** (1995), 478-488.

[7] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley and Sons Inc., New York (1992).

[8] 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.