AVERAGE COMPLEXITY OF ADDITIVE PROPERTIES FOR MULTIWAY TRIES : A UNIIFIED APPROACH

(Extended Abstract)

Wojciech Szpankowski*

Department of Computer Sciences
Purdue University
West Lafayette, IN 47907, U.S.A.

Abstract

We study multiway asymmetric tries. Our main interest is to investigate the depth of a leaf and the external path length, however we also formulate and solve a more general problem. We consider a class of properties called *additive properties*. This class is specified by a common recurrence relation. We give an exact solution of the recurrence, and present an asymptotic approximation. In particular, we derive all (factorial) moments of the depth of a leaf and the external path length. In addition, we solve an open problem of Paige and Tarjan about the average case complexity of the improved lexicographical sorting. These results extend previous analyses by Knuth [12], Flajolet and Sedgewick [6], Jacquet and Regnier [10], and Kirschenhofer and Prodinger [11].

1. INTRODUCTION

Digital searching is a well-known technique for storing and retrieving information using lexicographical (digital) structure of words. Let U be an alphabet containing V elements, and each element may occur with different probability (asymmetric trie). A *trie* or *radix search trie* is such a V-ary digital search tree that edges are labeled by elements from U and leaves (external nodes) contain the keys [1],[8], [12]. The access path from the root to a leaf is a minimal prefix of the information contained in the leaf. An important variant of tries is obtained using a sequential storage algorithm for subtries with the size less than or equal to a fixed bound b. In other words, each external node is capable of storing at most b keys. Such a trie will be called *b-trie* [3], [5].

This paper presents a thorough analysis of *b–tries* from the depth of a leaf point of view. Although we focus our attention on the depth and external path length, we consider them as motivating examples for general studies of the average complexity of so called *additive properties* for tries. Roughly speaking, a trie property is additive if its recurrence description is linear. For example, the depth of a leaf, the external path length, the number of internal nodes are additive properties, while the height of a trie is not. We discuss a class of additive properties which possess a common recurrence equation. The exact solution and asymptotic approximation of the recurrence are obtained. These results are then used in a number of applications. We present all factorial moments of the depth of a leaf and the external path length in a b- trie. In particular, we prove that the m-th factorial moment of the depth of insertion is $\alpha \ln^m n + \beta \ln^{m-1} n + O(\ln^{m-2} n)$, where α and β are some constants dependent on the distribution of elements in the alphabet. This implies that the variance of the depth is either equal to $\alpha \ln n + O(1)$ for asymmetric tries or only $O(1)$ for a symmetric trie. We also compute the average number of internal nodes and the number of internal nodes with all sons external nodes. The average external

* and Technical University of Gdansk, Poland.

path length and the average number of internal nodes are used to solve an open problem of Paige and Tarjan [17] about the average complexity of the improved lexicographical sorting.

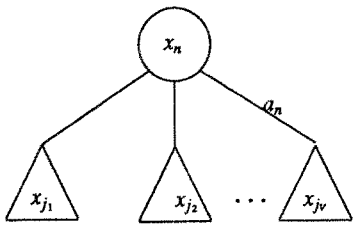
Finally, let us mention that the average case complexity for symmetric V-ary tries was discussed [12], [6]. Recently, Kirschenhofer and Prodinger [11] studied the variance of the depth of insertion for symmetric binary tries, while Jacquet and Regnier [10] obtained the limiting distribution for the depth of insertion for binary tries. This paper extends all of these results. We omit here all proofs, and provide them in the final version of the paper.

2. NOTATIONS AND MAIN RESULTS

Let us consider a set T_n of all *b–tries* with n keys over an alphabet $U = \{\sigma_1, \ldots, \sigma_V\}$. We assume that a key $x = \{x_1, x_2, \ldots, x_k, \ldots\}$ is a sequence of elements from U which form an independent sequence of Bernoulli trials. That is, for any k the probability $Pr\{x_k = \sigma_i\} = p_i$, $\sum_{i=1}^{V} p_i = 1$, and p_i does not depend on k. Such an approach is known as *Bernoulli model* [3], [5], [10]. In many applications such trie properties (parameters) as: *the depth of a leaf* (the number of internal nodes in the trie on the path from the root to the leaf), *the external path length* (the sum of the depths of all leaves), *the number of internal nodes, the height* (the maximum over all depths) , etc. are of particular interests. We shall study a class of properties called further *additive properties* which possess a common description through a recurrence equation. We focus our attention on the average case complexity of such properties.

2.1 Additive properties of b-tries

Let us consider an asymmetric b-trie with n records, and let Ξ be a property of such a trie. For example, Ξ might be the depth of a leaf , the external path length , the number of internal nodes, the height , etc. We denote by x_n the quantitative value of Ξ . Let a_n represent the value of the property Ξ that is possessed by the root of a trie. Then, in studying the average complexity of the property Ξ we may use the following graphical representation:



In other words, the value x_n of the property Ξ is a (recursive) function of the values $x_{j_1}, \ldots, x_{j_V}$ ($j_1 + \cdots + j_V = n$) of Ξ in all subtries of the trie, and the amount a_n of the property Ξ possessed by the root. If x_n is an additive function of $x_{j_1}, \ldots, x_{j_V}$, then the property Ξ is called additive. More precisely, we need that x_n satisfies the following recurrence:

$$\begin{aligned} &\textbf{given: } x_0, x_1, \ldots, x_B, \\ &\textbf{solve: } x_n = a_n + \sum_{j_\Sigma = n} \left[\begin{matrix} n \\ \mathbf{j} \end{matrix} \right] p_1^{j_1} \cdots p_V^{j_V} (x_{j_1} + \cdots + x_{j_V}), \quad n > B, \end{aligned} \tag{1}$$

where a_n is a given sequence (we also call it additive term) , $\sum_{i=1}^{V} p_i = 1$, and B is an integer . In (1) we have used the the following notations: $\left[\begin{matrix} n \\ \mathbf{j} \end{matrix} \right] = \left[\begin{matrix} n \\ j_1, \ldots, j_V \end{matrix} \right] = \dfrac{n!}{j_1! j_2! \ldots, j_V!}$ and $\sum_{j_\Sigma = n} f(j_1, \ldots, j_V)$ is a sum of $f(j_1, \ldots, j_V)$ over all $\mathbf{j}$ such that $j_1 + j_2 + \cdots + j_V = n$. We shall study properties Ξ for which their quantitative values x_n have the above description, and such properties are called additive. For example, the depth of a leaf, the external path length, the number of internal nodes are additive properties, while the height of a trie is not.

It turns out that a general solution of the recurrence depends on a transformed sequence, $\hat{a}_n$, of a_n defined as:

$$\hat{a}_n = \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] (-1)^k a_k.$$

The pair a_n and $\hat{a}_n$ is called *inverse relations* [15], since $\hat{\hat{a}}_n = a_n$. Let also

$$\beta_i = -\{x_i - a_i - \sum_{j_\Sigma = i} \left[\begin{matrix} i \\ \mathbf{j} \end{matrix} \right] p_1^{j_1} \cdots p_V^{j_V} \sum_{k=1}^{V} x_{j_k}\}/i!, \qquad i = 1,2, \ldots, B.$$

Note that the generating function of x_n, i.e., $X(z) = \sum_{n=0}^{\infty} x_n \dfrac{z^n}{n!}$, satisfies the following functional equation

$$X(z) - \sum_{i=1}^{V} X(p_i z) e^{(1-p_i)z} = A(z) - \sum_{i=0}^{B} z^i \beta_i$$

Define now $\Phi(z) = X(z) e^{-z}$, and transform the above into functional equation on $\Phi(z)$. Using Taylor expansion of $\Phi(z)$ one proves that

LEMMA 1. For any n, the recurrence (1) possesses the following solution

$$x_n = x_0 + n(x_1 - x_0) + \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{\hat{a}_k - \sum_{r=0}^{\min\{B,k\}} (-1)^r \left[\begin{matrix} k \\ r \end{matrix} \right] r! \beta_r}{1 - \sum_{i=1}^{V} p_i^k}. \tag{2}$$

Proof: See [14]. □

Solution (2) simplifies if $x_0 = x_1 = \cdots = x_B = 0$. In this case, in fact, we are also able to compute $\hat{x}_n$ which is necessary to obtain the m-th factorial moment of the depth of a leaf.

COROLLARY 1. If $x_0 = x_1 = \cdots = x_B = 0$, then

$$x_n = \sum_{k=2}^{n} (-1)^k \begin{bmatrix} n \\ k \end{bmatrix} \frac{\hat{a}_k - \sum_{r=0}^{B} (-1)^r \begin{bmatrix} k \\ r \end{bmatrix} a_r}{1 - \sum_{i=1}^{V} p_i^k}, \tag{3}$$

and

$$\hat{x}_n = \frac{\hat{a}_n - \sum_{r=0}^{B} (-1)^r \begin{bmatrix} k \\ r \end{bmatrix} a_r}{1 - \sum_{i=1}^{V} p_i^k}. \tag{4}$$

Proof: Eq (3) follows directly from (2), and (4) is a consequence of the definition of the inverse relations.

□

2.2 Applications of Lemma 1

Factorial moments of the depth of a leaf and the external path length

There is a simple relationship between the depth of insertion (depth of a leaf) and the external path length. Let $H_n(z)$

the external path length be the generating function with the coefficient at z^k representing the expected number of external nodes at level k in the family of tries T_n built from n records. Also let L_n and D_n be random variables representing the external path length and the depth of a leaf in the family T_n. The m-th factorial moment $d_n^{\underline{m}}$ of D_n is defined as :

$$d_n^{\underline{m}} = E\{D_n(D_n - 1) \cdots (D_n - m + 1)\}$$

In a similar way we define factorial moments of L_n . However, it is more convenient to introduce a *generalized external path length*. Let $D_n(i)$ denote the path length from the root to the i-th record in the family T_n, where $i = 1,2,...,n$. Then, the generalized external path length of order m is defined as

$$L_n^{\underline{m}} = \sum_{i=1}^{n} D_n(i)[D_n(i)-1]\cdots[D_n(i)-m+1]$$

and let $l_n^{\underline{m}} \overset{def}{=} EL_n^{\underline{m}}$. We call $l_n^{\underline{m}}$ the average of order m of the external path length $L_n^{\underline{m}}$. Note that $l_n^{\underline{m}}$ is *not* the m-th factorial moment of the external path length L_n, but the moments of L_n are simply related to $l_n^{\underline{m}}$, e.g., the average of L_n is equal to $l_n^{\underline{1}}$, and the variance of the external path length is $l_n^{\underline{2}} + l_n^{\underline{1}} - \frac{1}{n}(l_n^{\underline{1}})^2$. The following lemma establishes a relationship between $H_n(z)$ and the moments.

LEMMA 2. For any natural m and n the following holds

$$H_n(1) = n, \quad \frac{d^m H_n(z)}{dz^m}\Big|_{z=1} = H^{(m)}(1) = l_n^{\underline{m}},$$

$$d_n^m = l_n^m / n.$$

Proof. It follows directly from the above definitions.

□

There is no explicit formula for $H_n(z)$, but a rather nice recurrence.

LEMMA 3. For any natural n and b, $H_n(z)$ satisfies the following recurrence

$$H_n(z) = n \quad \text{for} \quad n \le b$$

$$H_n(z) = z \sum_{j_z = n} \left[\begin{matrix} n \\ \mathbf{j} \end{matrix} \right] p_1^{j_1} p_2^{j_2} \cdots p_V^{j_V} [H_{j_1}(z) + \cdots + H_{j_V}(z)] \quad \text{for} \quad n > b.$$

Proof. See Knuth [12].

□

By Lemma 2 to compute all factorial moments of the depth of insertion we need l_n^m. Using Lemma 3 we may prove that l_n^m $m = 1,2, \ldots,$ satisfy a system of recurrence equations.

THEOREM 1. The average of order m of L_n^m is given by the following recurrence

$$l_n^m = 0 \quad \text{for} \quad n \le b$$

$$l_n^m = m! \sum_{i=1}^{m} (-1)^{m-i} \frac{l_n^{i-1}}{(i-1)!} + \sum_{j_z = n} \left[\begin{matrix} n \\ \mathbf{j} \end{matrix} \right] p_1^{j_1} \cdots p_V^{j_V} (l_{j_1}^m + \cdots + l_{j_V}^m) \quad n > b,$$

where $l_n^0 = n$.

Proof. The proof is by induction, and it is left for the reader.

□

By Theorem 1 l_n^m satisfies (1) with $a_n = m! \sum_{i=1}^{m} (-1)^{m-i} \frac{l_n^{i-1}}{(i-1)!}$. For example, for m=1 one immediately obtain from Lemma 1 and Theorem 1

$$l_n^1 = \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{\sum_{r=1}^{b} (-1)^{r+1} \left[\begin{matrix} k \\ r \end{matrix} \right] r}{1 - \sum_{i=1}^{V} p_i^k},$$

and

$$l_n^1 = \sum_{r=1}^{b} (-1)^{r+1} \left[\begin{matrix} k \\ r \end{matrix} \right] \frac{r}{1 - \sum_{i=1}^{V} p_i^k}.$$

Let now $m = 2$ Then by Theorem 1 $a_n^{(2)}=2[\,l_n^{\underline{1}} - n\,]$, and $\hat{a}_n^{(2)} = 2\hat{l}_n^{\underline{1}} - 2\delta_{n,1}$ [15]. After some algebra, we find that

$$l_n^{\underline{2}} = 2 \sum_{r=1}^{b} (-1)^{r+1}\, r \sum_{k=2}^{n} (-1)^n \begin{Bmatrix} n \\ k \end{Bmatrix} \begin{bmatrix} k \\ r \end{bmatrix} \frac{\sum_{i=1}^{V} p_i^k}{(1 - \sum_{i=1}^{V} p_i^k)^2}.$$

Generalizing the above, and applying recursively Lemma 1 , Corollary 1 and Theorem 1 we obtain the exact solution for the m-th factorial moment of the depth.

PROPOSITION 1. For all n the moment of order m, $l_n^{\underline{m}}$, of the generalized external path length in a b-trie is given by

$$l_n^{\underline{m}} = m! \sum_{r=1}^{b} (-1)^{r+1}\, r \sum_{k=2}^{n} (-1)^k \begin{Bmatrix} n \\ k \end{Bmatrix} \begin{bmatrix} k \\ r \end{bmatrix} \frac{[\,\sum_{i=1}^{V} p_i^k\,]^{m-1}}{[\,1 - \sum_{i=1}^{V} p_k^k\,]^m}, \quad n > b, \tag{5}$$

and $l_n^{\underline{m}} = 0$ for $n \le b$. The m-th factorial moment of the depth, $d_n^{\underline{m}}$ is $l_n^{\underline{m}}/n$.

□

The average number of internal nodes and other applications

A number of other applications of Lemma 1 is possible. For example, from the storage view point it is important to know the average number of internal nodes, I_n. Naturally, I_n is an additive property, hence the recurrence (1) is satisfied with $x_0 = \cdots x_b = 0$, and $a_n = 1$. Applying Corollary 1 we obtain immediately

$$I_n = \sum_{k=2}^{n} (-1)^k \begin{Bmatrix} n \\ k \end{Bmatrix} \frac{\sum_{r=0}^{b} (-1)^{r+1} \begin{bmatrix} k \\ r \end{bmatrix}}{1 - \sum_{i=1}^{V} p_i^k}$$

In some other applications we might be interested in the average number of internal nodes with all sons external nodes (see [12], [6]). Assume for simplicity that $b=1$ and $V=2$. Let E_n denote the average number of such nodes. Then $E_0=E_1=0$, $E_2=1$ and for $n>2$ the average E_n satisfies the recurrence (1) with $a_n=0$. Note that $\beta_0=\beta_1=0$ and $\beta_2=1-p_1^2-p_2^2$. Hence by Lemma 1 and straightforward computations we find

$$E_n = \sum_{k=2}^{n} (-1)^k \begin{Bmatrix} n \\ k \end{Bmatrix} \begin{bmatrix} k \\ 2 \end{bmatrix} \frac{1-p_1^2-p_2^2}{1-p_1^k-p_2^k}$$

The generalization for $V>2$ is simple but need some additional computations. For other applications of (1) see [3] - [8], [13], [14].

Asymptotic approximation

From the practical view point it is important to know asymptotic approximation of $d_n^{\underline{m}}, I_n, E_n$, etc. However, instead of computing the approximation for each of the above quantities we may equivalently determine the asymptotic approximation of the following:

$$S(n,r,m) \stackrel{def}{=} \sum_{k=2}^{n} (-1)^k \begin{Bmatrix} n \\ k \end{Bmatrix} \begin{bmatrix} k \\ r \end{bmatrix} \frac{\alpha^k}{(1 - \sum_{i=1}^{V} p_k^k)^m}, \tag{6a}$$

where $r = 1,2,\ldots, B$ and α is a parameter. (Note that $S(n,r,m)$ is the sum in (2) if $a_n = \begin{bmatrix} n \\ r \end{bmatrix} \alpha^n$). The application of (6a) to the evaluation of $l_n^{\underline{m}}$, I_n and E_n is straightforward. For example, the m-th factorial moment, $l_n^{\underline{m}}$, is expressed in terms of $S(n,r,m)$ as follows

$$l_n^{\underline{m}} = m! \sum_{j=m-1} \begin{bmatrix} m-1 \\ j \end{bmatrix} \sum_{r=1}^{b} (-1)^{r+1} r\, S(n,r,m)$$

where $\alpha = \prod_{s=1}^{V} p_s^{j_s}$.

To evaluate (6a) we may use either Rice's method [6], [7] or Mellin transform technique [6], [7], [9], [13], [16]. We apply here the latter method. We proved in [14] that $S(n+r,r,m) = T(n+r,r,m) + O(1)$, where

$$T(n+r),r) = (-1)^r \frac{n+r}{r!} \alpha[1+O(n^{-1})] \int_{(\frac{1}{2} - [2-r]^+)} \frac{\Gamma(z)(n\alpha)^{r-1-z}}{(1 - \sum_{i=1}^{V} p_i^{r-z})^m}\, dz, \tag{6b}$$

and $\Gamma(z)$ is the gamma function [16], $a^+ = \min\{0,a\}$ and the integral notation $\int_{(c)} f(\cdot)$ stands for $\frac{1}{2\pi_i} \int_{c-i\infty}^{c+i\infty} f(\cdot)$. The line of integration is either $(-3/2 - i\infty, -3/2 + i\infty)$ for $r=0$, or $(-1/2 - i\infty, -1/2 + i\infty)$ for $r=1$ or $(1/2 - i\infty, 1/2 + i\infty)$ for $r>1$.

The evaluation of the counter integral in (6b) is routine: one goes from $(c, -iN_1)$ to (c, iN_1) to (N_2, iN_1) to $(N_2, -iN_1)$ to $(c, -iN_1)$ in a negative sense, where $c = \frac{1}{2} - [2-r]^+$. For $N_1 \to \infty$ the horizontal parts of the integral vanish since $\Gamma(t + iN_1) = O(1 + iN_1|^{t-\frac{1}{2}} e^{-t - \pi N_1/2})$ [16], while the vertical component over $(N_2, -iN_1)$ decays due to the factor n^{r-1-z} [12], [16]. Hence the required integral is minus the sum of residuals of the function under the integral to the right of the vertical line fixed at point $c = \frac{1}{2} - [2-r]]^+$. The details may be found in [14].

For $m = 1$ a closed form expression for $S(n,r,1)$ as n tends to infinity is available. Let us define

$$h_n = (-1)^n \sum_{i=1}^{V} p_i\, ln^n\, p_i, \quad n > 0.$$

and $h_0 = 0$. Then

PROPOSITION 2. For any n and r the following holds

$$S(n,0,1)) = n\,\alpha\{\frac{ln(n\alpha)+\gamma-1}{h_1} + \frac{h_2}{2h_1^2} + f(n\alpha)\} + O(1)$$

$$S(n,1,1) = n\alpha\{\frac{\ln(n-1)\alpha+\gamma}{h_1} + \frac{h_2}{2h_1^2} - f((n-1)\alpha)\} + O(1)$$

$$S(n,r,1) = (-1)^r\,\frac{n\alpha}{r!}\{\frac{(r-2)!}{h_1} + f((n-r)\alpha)\} + O(1) \quad r>2$$

where γ=0.573 is the Euler constant, and $f(n)$ is a fluctuating function with a very small amplitude [12], [8], [6]. (In practise, the function $f(n)$ may be savely ignored).

Proof: The proof may be found in [13].

□

In particular, using the above we immediately obtain an asymptotic approximation for the average number of internal nodes, I_n. We find that

$$I_n = \frac{n}{h_1}\{[1-\sum_{r=2}^{b}\frac{1}{r(r-1)}]+f(n)\}+O(1)$$

On the other hand, the average number of nodes with both sons external nodes is given by

$$E_n = n\,\frac{1-p_1^2-p_2^2}{2h_1}(1+f(n))+O(1)$$

To evaluate d_n^m for large n we need the asymptotic approximation of $S(n,r,m)$ for $m>1$. This is more difficult, however, we can prove that

PROPOSITION 3. For any m, and n large enough

$$d_n^m = \frac{1}{h_1^m}\,ln^m n + \frac{m}{h_1^m}\,ln^{m-1}n[\gamma + \frac{m}{2}\,\frac{h_2}{h_1} - (m-1)h_1 - H_{b-1} - h_1^m\,F(n)] + O(ln^{m-2}n)$$

where H_{b-1} is the $(b-1)$-st harmonic number [12], and $F(n)$ is a fluctuating function with a very small amplitude.

Proof: We use extensively (6b). Algebra may be found in [14].

□

Two moments play usually an important role in tries analysis, namely: the average and the variance, σ_n^2, of the depth of insertion. Using the above approach we obtain immediately

PROPOSITION 4. (i) The average depth of a leaf is given by

$$d_n = \frac{1}{h_1} \ln n + \frac{1}{h_1}[\gamma + \frac{h_2}{2h_1} - H_{b-1}] - F(n) + O(n^{-1}). \tag{7}$$

(ii) The variance, σ_n^2, of the depth is

$$\sigma_n^2 = \frac{h_2 - h_1^2}{h_1^3} \ln n + C + F(n) + O(n^{-1}), \tag{8}$$

where

$$C = 2\alpha - \frac{2\gamma}{h_1} - \frac{h_2}{h_1^2} + \frac{2H_{b-1}}{h_1}(1 - \frac{h_2}{h_1^2} - \frac{\gamma}{h_1}) + \frac{2e_{b-1}}{h_1^2} +$$
$$(\frac{\gamma}{h_1} + \frac{h_2}{2h_1^2} - \frac{H_{b-1}}{h_1})(1 - \frac{\gamma}{h_1} - \frac{h_2}{2h_1^2} + \frac{H_{b-1}}{h_1})$$

and

$$\alpha = \frac{1}{h_1^2}[\frac{\pi^2}{12} + \frac{\gamma^2}{2} + \frac{3h_2^2}{4h_1^2} + \frac{\gamma h_2}{h_1} - \frac{h_3}{3h_1}],$$

and e_b is defined as $e_b = \sum_{r=1}^{b} \frac{H_{r-1}}{r}$ ($H_0=0$, and $e_0=0$). $F(n)$ is a fluctuating function with a small amplitude. In particular, for V-ary symmetric tries $h_n = \ln^n V$ hence (8) implies

$$\sigma_n^2 = \frac{\pi^2}{6\ln^2 V} + \frac{1}{12} - \frac{H_{b-1}^2}{\ln^2 V} + \frac{2e_{b-1}}{\ln^2 V} + F(n) + O(n^{-1}) \tag{9}$$

(iii) The variance of the external path length $\sigma_{L_n}^2$ is equal to

$$\sigma_{L_n}^2 = n\,\sigma_n^2$$

so it is $O(n \ln n)$ for asymmetric tries, and $O(n)$ for symmetric tries.

Proof: Equation (7) follows immediately from $l_n^1 = \sum_{r=1}^{B} (-1)^{r+1} r\, S(n,r,1)$, and Proposition 2. To compute σ_n^2 note that $\sigma_n^2 = l_n^2 + l_n^1 - [l_n^1]^2/n$, and $l_n^2 = \sum_{r=1}^{b} (-1)^{r+1} r\, S(n,r,2)$.

□

The table below shows the variance of the depth, σ_n^2, for symmetric V-ary b-tries. (see (9)). Note that by (7) the influence of b on the average depth is of order $O(1)$, and for small values of b it may be safely ignored in practise. However, the variance critically depends on b, and for larger b we obtain more balanced tries. For example, for V=2 the variance σ_n^2 decreases from 3.507 for b=1 to 0.6741 for b=4. But there is a trade-off between b and the average searching time. Note that bigger b implies larger searching time in the sequential file of the external node. The total average searching time is $d_n + (b-1)/2$, where d_n is given by (7). Hence, the searching time is minimized for b optimal equal to $b_{opt} = 1/(2h_1) + 1$.

Table. Variance of the depth for symmetric b-tries

	σ_n^2			
V	$b=1$	$b=2$	$b=3$	$b=4$
2	3.5070	1.4256	0.9053	0.6741
3	1.4462	0.6177	0.4105	0.3184
4	0.9393	0.4189	0.2888	0.2310
5	0.7183	0.3323	0.2358	0.1929
6	0.5957	0.2842	0.2063	0.1717

3. APPLICATIONS AND DISCUSSION

In this section we show some of the possible applications of the above results. In particular, we offer the average complexity of the improved lexicographical sorting algorithm proposed recently by Paige and tarjan [17].

Optimization problems.

Let us consider d_n^m as a function of $\mathbf{p} = (p_1, p_2, \ldots, p_V)$. Then a question arises what is an optimal choice of $\mathbf{p}$? It is intuitively clear that the average depth of insertion is minimized for the symmetric case. However, using Proposition 1 it is easy to notice that l_n^m and d_n^m are minimized for *all* n and m if the trie is a symmetrical one, that is, $p_1 = p_2 = \cdots = p_V = 1/V$. Naturally, the bigger the V is, the smaller the average depth of insertion is, however, the data structure becomes more complicated. Moreover, formula (7) shows that the bigger the b is, the smaller the average depth of insertion is, however, the impact of b is of the secondary importance since the leading factor in (7) does not depend on b.

A measure of balance for a tree.

The variance of the depth of insertion might be considered as a measure of *how well a tree is balanced.* In the height-balanced trees the depth of a leaf is the same (or almost the same) for all leaves. Then, the variance of the depth is equal to zero. For other trees the depth is a random variable, however, the smaller the variance is, the more balanced the tree is. Indeed, by Tchebyshev inequality, we know that $Pr\{|D_n - d_n| > \delta\} < \sigma_n^2/\delta^2$. For example, let $\delta = 3\sigma$, then $Pr\{|D_n - d_n| > 3\sigma_n\} < 1/9$, and it says that with probability 0.11 the depth lies in the interval $(d_n - 3\sigma_n, d_n + 3\sigma_n)$, hence the smaller σ is, the smaller the interval is. This also means that for small σ the average of the depth of insertion is a good measure of the actual depth, while for larger σ, it is very poor performance issue. Let us apply this to tries. By (8) we see that for symmetric tries $h_2 - h_1^2 = 0$, hence $\sigma_n^2 = O(1)$ and does not depend on n. We may claim that symmetric tries are of an order of magnitude better balanced than asymmetric tries. Let $V = 2$, then for $p = 0.5$ (symmetric trie) $\sigma_n^2 \approx 3.507$, while for $p = 0.1$ $\sigma_n^2 = 12.64 \ln n + O(1)$ and for $p = 0.3$ $\sigma_n^2 = 0.66 \ln n + O(1)$. The Tchebyshev inequality implies that with probability 0.11 the depth of insertion for a symmetric trie with $V = 2$ lies in $(d_n - 5.5, d_n + 5.5)$, while with the same probability the depth is in the interval $(d_n - 10\sqrt{\ln n}, d_n + 10\sqrt{\ln n})$ for $p = 0.1$ and in

$(d_n - 2.4\sqrt{\ln n}\,,\ d_n + 2.4\sqrt{\ln n}\,)$ for $p = 0.3$ and large n. Note also that bigger the b is, more balanced the trie is.

Improved lexicographical sorting

Paige and Tarjan proposed an improved lexicographical sorting algorithm [17]. It works in two steps. The first determines so called *significant prefix* by building a trie over an alphabet, assuming that the total length of all strings is equal to L. The proposed algorithm runs in $O(L')$ time where L' is the total length of all significant prefixes. Aho, Hopcroft and Ullman [1] gave a solution with $O(L)$ worst case asymptotic time. Hence, the ration L/L' indicates the improvements over the Aho et al algorithm.

To compute the average complexity of the improved lexicographical sorting, and compare it with the Aho's lexicographical sorting, we first introduce some notations. Let $S = \{x_1, x_2, \ldots, x_n\}$ be set of the *finite* length strings built randomly over an V-ary alphabet U subject to the total length of all strings L (L is fixed). Let also L' denote the total length of all significant prefixes. Note that L' is a random variable, and it is equivalent to the external path length in the appropriate trie. Assuming symmetric alphabet by Proposition 4 we find that the average value of the external path length is $n\ lg_V\ n + O(1)$. Note, however, that this does not follow directly from our previous results, since in our model we have assumed unlimited strings. Nevertheless, it is easy to show that for large n, and distinct keys (strings), the above holds. Hence $EL' = n\ lg_V\ n + O(1)$, and the improved ratio $IMP \stackrel{def}{=} L/EL' = L/[n\ lg_V\ n + O(1)]$. Such a formula is not very informative, since there is a relationship between L and n. Indeed, subject to L the number of strings, n, might be equal to one, or two or ... or $n_{\max}$, where $n_{\max}$ is the maximum number of finite strings whose total length is L. Naturally, $n \le n_{\max}$, and if $n = n_{\max}$ (the trie in that case is almost a complete V-ary tree), then $L = n\ lg_V\ n + O(1)$, hence $IMP \simeq 1$ and the improved algorithm runs the same time as the Aho's lexicographical algorithms. The improvement depends on the relationship between n and $n_{\max}$.

Let us first compute $n_{\max}$. We obtain the maximum number of strings packed into L if we take V strings of length one, V^2 strings of length two, ..., V^κ strings of length κ, where κ is such an integer that

$$\sum_{l=1}^{\kappa} lV^l \le L \le \sum_{l=1}^{\kappa+1} lV^l. \tag{10}$$

Then, the maximum number of strings packed into L, $n_{\max}$, satisfies

$$\sum_{l=1}^{\kappa} V^l \le n_{\max} \le \sum_{l=1}^{\kappa+1} V^l. \tag{11}$$

Note, that by geometric series formula, (10) and (11) are equivalent to

$$V\ \frac{\kappa V^{\kappa+1} - (\kappa+1)V^\kappa + 1}{(V-1)^2} \le L \le V\ \frac{(\kappa+1)V^{\kappa+2} - (\kappa+2)V^{\kappa+1} + 1}{(V-1)^2} \tag{12}$$

$$V\ \frac{V^\kappa - 1}{V-1} \le n_{\max} \le V\ \frac{V^{\kappa+1} - 1}{V-1}. \tag{13}$$

But (12) and (13) imply that

$$\kappa n_{\max} + \frac{1}{V-1}(\kappa - n_{\max}) \le L \le (\kappa+1)n_{\max} + \frac{1}{V-1}(\kappa + 1 - n_{\max}).$$

For, by (13)

$$\kappa = lg_V\ n_{\max} - \frac{1}{V \ln V} + O(n_{\max}^{-1}),$$

and finally we obtain

$$L = n_{\max}\ lg_V - n_{\max}\left(\frac{1}{V \ln V} + \frac{1}{V-1}\right) + O(lg_V\ n_{\max}). \tag{14}$$

Let now $n_{\max} = \beta n$, $\beta \ge 1$. Then, the ratio of improvement, $IMP = L/EL'$ is given by

$$IMP = L/EL' = \beta\left[1 + \frac{lg_V \beta - 1/(V \ln V) - 1/(V-1)}{lg_V\ n}\right] + O(n^{-1}). \tag{15}$$

Note that for $\beta = 1$ (15) implies that $IMP \approx 1$ as expected.

4. CONCLUSIONS

In this paper we propose a unified approach to study additive properties of digital tries. This goal was achieved through a solution of a general class of linear recurrences. In particular, we derived all (factorial) moments of the depth of a leaf, we computed the average number of interval nodes, and the average number of nodes with both sons external nodes. Finally, we gave the average complexity of the improved lexicographical sorting algorithm proposed recently by Paige and Tarjan.

ACKNOWLEDGMENT

I wish to thank Professors Helmut Prodinger and Peter Kirschenhofer from the Technical University of Vienna for pointing out an error in the computation of the constant C in (8).

REFERENCES

[1] Aho, A., Hopcroft, J. and Ullman, J., Data structures and algorithms, Addison-Wesley, 1983.

[2] Frdelyi, A., Higher transcendental functions, McGraw-Hill Book, 1953.

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

[4] Fayolle, Ph., Flajolet, Ph, Hofri, M. and Jacquet, Ph., Analysis of a stack algorithm for random multiple-access communication, *IEEE Trans. on Information Theory*, **IT-31, 2,** 1985, 244-254.

[5] Flajolet, Ph., On the performance evaluation of extendible hashing and trie searching, *Acta Informatica* 20, 1983, 345-369.

[6] Flajolet, Ph. and Sedgewick R., Digital search trees revisited, *SIAM J. Compt.* ,15, 1986, pp. 748-767.

[7] Flajolet, Ph., Mathematical methods in the analysis of algorithms and data structures, INRIA Technical Report, 400, 1985.

[8] Gonnet G., Handbook of algorithms and data structures, Addison-Wesley, 1984.

[9] Henrici, P., Applied and computational complex analysis, John Wiley & Sons, New York 1977.

[10] Jacquet, Ph. and Regnier, M., Limiting distributions for trie parameters, Proc. of *CAAP'86*, 1985.

[11] Kirschenhofer, P. and Prodinger, H., Some further results on digital trees, *ICALP 86*, to appear.

[12] Knuth, D., The art of computer programming, Addison-Wesley, 1973.

[13] Szpankowski, W., Analysis of a recurrence equation arising in stack-type algorithms for collision-detecting channels, Proc. Intern. seminar on Computer Networking and Performance Evaluation, Tokyo 1985, 9-3-1-9-3-12.

[14] Szpankowski, W., Some results on the V-ary asymmetric tries, Purdue University, CDS-TR-582, 1985.

[15] Riordan, J., Combinatorial identities, John Wiley & Sons, 1968.

[16] Whittaker, E. and Watson, G., A course of modern analysis, Cambridge Press, 1935.

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