1986

On the Analysis of the Average Height of a Digital Trie: Another Approach

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:

86-646

ON THE ANALYSIS OF THE AVERAGE HEIGHT OF A DIGITAL TRIE: ANOTHER APPROACH

Wojciech Szpankowski

CSD-TR-646
December 1986

ON THE ANALYSIS OF THE AVERAGE HEIGHT OF A DIGITAL TRIE: ANOTHER APPROACH

*Wojciech Szpankowski**
Department of Computer Sciences
Purdue University
West Lafayette, IN 47907

Abstract

The average height of a digital trie has been recently investigated in many papers [2]–[8]. In most works on binary digital tries, a Bernoulli model and independent keys are assumed. We relax these assumptions in that V-ary asymmetric tries, Bernoulli and Poisson models, and dependent keys are considered. We show that the average height of the trie is asymptotically equal to 2 $lg_u\ n$ (for the Bernoulli model) and 2 $lg_u\ \mu$ (the Poisson model) where n and μ are the number of records and the average number of records respectively. The parameter u is defined as $u^{-1} = \sum_{i=1}^{V} p_i^2$, and the V elements of the alphabet are distributed according to probabilities p_i, $i = 1, \ldots, V$. Finally, a generalization to the so called b-tries is discussed. In contrast to the previous analysis, our approach is very simple since we avoid explicit computation of the height distribution.

* and the Technical University of Gdansk, Poland

1. INTRODUCTION

Let A be a V-ary alphabet, i.e., $A = \{\alpha_1 , \ldots , \alpha_V\}$ and let S denote the set of n strings (keys) built over the alphabet A. A *trie* (digital search time) is a V-ary digital search tree in which edges are labeled by elements from A and leaves (external nodes) contain the keys (records) [1]. 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 sequential storage algorithm for subtries with the size less than or equal to a fixed bound b, i.e. external node is capable of storing at most b keys. Such a trie is called b-trie [2], [3].

Digital tries find many applications in computer science. A trie is used as an index to access data in a secondary memory (e.g. extendible hashing) [2], [7], [8], [18], it can be used in the pattern-matching algorithms (position trees and string identifiers) [1] and in sorting algorithms like triesort [4], [16] and radix exchange sort [16],[19]. Some other applications of the digital tries include: conflict resolution algorithms for broadcast communications, polynomial factorization and Huffman's algorithm [1], [3], [16], [17], [19]. We analyze a random family of tries with n stored records from the height view point. It is assumed that each key consists of (possible infinite) elements from the alphabet A, and the element $\alpha_k \in A$, $k = 1, 2 , \ldots , V$, occurs with probability p_k at any position of a key (asymmetric V-ary trie). In most analyses (see [2]–[4], [7], [8]) binary symmetric tries were investigated which restricts the applications of the analysis (e.g., see matching-string problem where English characters occur with very different probabilities).

This paper provides a new methodology to study the average height of general asymmetric digital tries. Using a simple inequality for *order statistics* we prove that the average height EH_n, of a trie is $EH_n \sim 2lg_u \ n$, where $u^{-1} = \sum_{i=1}^{V} p_i^2$. This result is generalized in three different directions. At first, we drop the assumption that the fixed number of keys are stored in the trie.

Assuming that a trie is built over random number of keys distributed according to Poisson process with parameter μ (Poisson model), we prove that $EH_\mu \sim 2lg_u\, \mu$. Secondly, for b-tries we show that the average height is asymptocially equal to $(1 + \frac{1}{b})lg_u\, n$, where $u^{-1} = \sum_{i=1}^{V} p_i^{b+1}$. Finally, we assume that there exists some statistical dependency between keys. Then, it is proved that $EH_n = O(lg_u\, n)$, where u is a constant which reflex statistical dependency among the keys.

The average height of digital tries has been recently investigated in [2]–[8]. In [2] Flajolet studied binary symmetric b-tries. Based on some classical counting results in occupancy problems, Flajolet derived asymptotic distribution of the height. Using complex analysis (Cauchy integral formula) he also found the average height of a trie. Jacquet and Regnier [3] extended Flajolet's result to binary asymmetric tries. They have made extensive use of the Mellin transform technique. Devroye [4] analyzed binary symmetric tries, and based on the occupancy problem he derived some inequalities on the asymptotic distribution of the height. The most general results were obtained by Pittel [5] (see also [6]), where V-ary asymmetric tries with $b = 1$ were investigated. Unfortunately, the proofs in [5] and [6] are not constructive, and the results are well hidden. For some more results, see also [7] and [8]. Our approach to the problem is essentially different. In contrast to the previous analysis we use elementary calculus, and we avoid explicit computation of the height distribution. In this paper we only concentrate on the asymptotic results for the average height of digital tries, however, the methodology can be extended to the analysis of digital trees and Patricia tries.

2. MAIN RESULTS

Let us consider a set of all digital tries with n records, $X_1, X_2, \ldots, X_n$, over an alphabet $A = \{\alpha_1, \alpha_2, \ldots, \alpha_v\}$. Each record consists of (possible infinite) string of elements (digits) from A, e.g., $X_k = (x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots)$ where $x_{kj} \in A$, $j = 1, 2, \ldots$. For a given keys $X_1, X_2, \ldots, X_n$ the digital trie is built in a usual manner (see [1]). For example, in Figure

1 we show a 3-ary trie built over $A = \{1, 2, 3\}$ with 6 records $A, B, \ldots, F$. Note that a trie consists of two types of nodes, namely internal nodes and external nodes. The internal nodes are used to determine branching strategy, while keys (records) are stored in the external nodes.

The common assumptions under which the random family of tries is analyzed, are specified below:

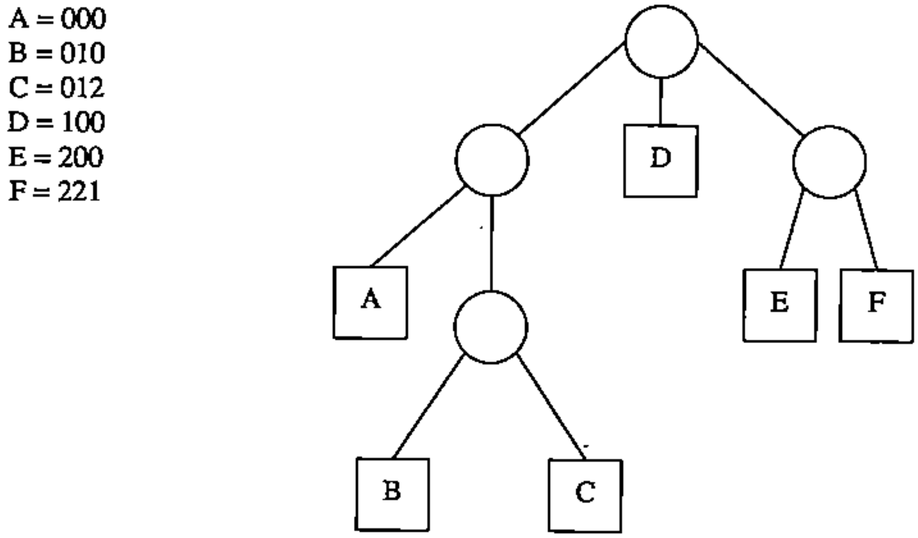

Figure 1. Example of 3-ary digital trie with n=6.

(i) A key $X_k = (x_{k1}, x_{k2} \ldots)$ is a sequence of elements (digits) from A which form an independent sequence of Bernoulli trials with $Pr\{x_{kj} = \alpha_i\} = p_i$, $k = 1, 2, \ldots, n$, $i = 1, 2, \ldots, V$.

(ii) The keys $X_1, X_2, \ldots, X_n$ are statistically independent.

(iii) The number of records stored in a trie is fixed and equal to n.

These assumptions create the so called *Bernoulli model.* In addition, we also assume that

(iv) the external node is capable to store only one record, i.e., regular tries ($b = 1$) are analyzed in this section.

In a trie three quantities are of particular interests: the depth of a leaf (the paths from the root to a randomly chosen leaf), the height, H_n, of a trie (the maximum over all depths), and the smallest path from the root to a leaf. The depth of a leaf was previously analyzed in [3], [5], [6] and [9]. Here we concentrate on the average height, EH_n.

Let us define a common path of two keys, com$(X_{i,} X_j)$, $i, j = 1, 2, \ldots, n$, as the common prefix of X_i and X_j, that is, com$(X_i, X_j) = k$ if X_i and X_j agree exactly on their first k digits, but differ in their $(k+1)$-st. Let $Y_{ij} = \mathrm{com}(X_i, X_j)$, $i \neq j$. Note that $Y_{ij} = Y_{ji}$, hence we restrict the indices to $i = 1, 2, \ldots, n$, $j = i+1, i+2, \ldots, n$. Sometimes, for simplicity, we renumber the random variables Y_{ij}, and we write $Y_1, Y_2, \ldots, Y_m$ where $m = n(n-1)/2$. There is, of course, a one-to-one correspondence between Y_{ij} and Y_k. Under the assumptions (i)–(iv) the random variable Y_{ij}, for any i and j, is geometrically distributed with parameter $p_1^2 + p_2^2 + \cdots + p_V^2$, that is

$$Pr\{Y_{ij} = k\} = Pr\{\mathrm{com}(X_i, X_j) = k\} = \left[\sum_{l=1}^{V} p_l^2\right]^k \left[1 - \sum_{l=1}^{V} p_l^2\right] \qquad k = 0, 1, \ldots \quad . \tag{1}$$

Let $u^{-1} \overset{def}{=} \sum_{l=1}^{V} p_l^2$. Note that although X_i, $i = 1, \ldots, n$ are independent, the random variables Y_{ij} are dependent.

To find a relationship between the height H_n and com(X_i, X_j) note that the common prefix of a particular key X_k and all other keys X_j, $j = 1, 2, \ldots, n$, $j \neq k$, determine the position of X_k in the trie. Hence

$$H_n = 1 + \max_{1 \le i \le n} \max_{j \ge i}\{Y_{ij}\} = 1 + \min_{1 \le k \le m}\{Y_k\}. \tag{2}$$

To illustrate (2), let us consider the trie in Figure 1. We find that $H_6 = 3$, and com$(A, B) = \mathrm{com}(A, C) = 1$, com$(A, D) = \mathrm{com}(A, E) = \mathrm{com}(A, F) = 0$; com$(B, C) = 2$, com$(B, D) = 0$, etc. Hence $3 = H_6 = 1 + \max\{\mathrm{com}(X_i, X_j)\} = 1 + \mathrm{com}(D, C) = 3$.

Eq.(2) suggests that to compute H_n we need to know some statistics of the maximum of m dependent random variables, $Y_1, Y_2, \ldots, Y_m$. Such a statistic is known in the literature as the *order statistic*. In the next subsection, we derive some simple properties of the average of $\max\{Y_k\}$, that is, $E\ \max\{Y_k\}$.

The average value of $\max\{Y_i\}$

Let $Y_1, Y_2, \ldots, Y_m$ be identically distributed random variables with the distribution function $F(y)$. Define

$$M_m = \max_{1 \le i \le m} \{Y_i\}.$$

It is easy to see that

$$M_m \le a_m + \sum_{i=1}^{m} (Y_i - a_m)^+ \tag{3}$$

where a_m is a parameter dependent on m, and $x^+ = \max\{0, x\}$. For a nonnegative random variable Y with distribution function $F(y)$ the average EY may be computed as $EY = \int_0^\infty [1 - F(y)]dy$. Hence, by (3) the average EM_m is

- *for continuous random variables*

$$EM_m \le a_m + m \int_{a_m}^{\infty} [1 - F(x)]dx \tag{4a}$$

- *for discrete random variables*

$$EM_n \le a_m + m \sum_{k=a_m}^{\infty} [1 - F(k)] \tag{4b}$$

The RHS of (4) is minimized if a_m is chosen such that

$$a_m = \min\{k : Pr\{Y > k\} \le \frac{1}{m}\}. \tag{5}$$

EXAMPLE 1: *Exponential distribution*

Let $F(y) = 1 - e^{-\lambda y}$, λ is a parameter. Then by (5) $a_m = \frac{1}{\lambda}\, ln\ m$, and (4a) implies

$$EM_m \le \frac{1}{\lambda}\, ln\ m + \frac{1}{\lambda}. \tag{6}$$

If, in addition, $Y_1, \ldots, Y_m$ are independent, then [10]

$$EM_m = \frac{1}{\lambda}\, ln\ m + \frac{\gamma}{\lambda} \tag{7}$$

where $\gamma = 0.577$ is the Euler constant. Note that the difference between (6) and (7) is of order $O(1)$.

□

EXAMPLE 2: *Geometric distribution*

Let Y be geometrically distributed, i.e. $Pr\{Y = k\} = p^k(1-p)$. Then $Pr\{Y > k\} = p^{k+1}$, and by (5) $a_m = \lfloor \frac{ln\ m}{ln\ p^{-1}} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor operator. Also by (4b)

$$EM_m \le \frac{ln\ m}{ln\ p^{-1}} + \frac{p}{1-p}. \tag{8}$$

Note that the geometric distribution may be approximated by an exponential distribution with parameter $\lambda = ln\ p^{-1}$. Since $ln\ p^{-1} \le \frac{1-p}{p}$ one finds that (8) is equivalent to (6) with $\lambda = ln\ p^{-1}$.

□

Both inequalities (6) and (8) imply that the leading term in EM_m is a_m. The question is whether $EM_m \sim a_m$, i.e. $\lim_{m \to \infty} EM_m / a_m = 1$. Lai and Robbin proved [10], [11] that $EM_m \sim a_m$ if the distribution $F(y)$ satisfies the following conditions

$$\lim_{y \to \infty} \frac{1 - F(cy)}{1 - F(y)} = 0 \quad \text{for every} \quad c > 1 \tag{9a}$$

$$\int_{-\infty}^{0} |x|^r \, d\, F(x) < \infty \text{ for some } r > 0. \tag{9b}$$

Note that (9) holds for the exponential and geometric distributions, i.e., $EM_m \sim \frac{1}{\lambda}\, ln\ m$.

The average height of a trie

By (2) $EH_n = 1 + E \max_{1 \le k \le m} \{Y_k\}$, where $m = n(n-1)/2$ and Y_k is geometrically distributed with parameter $u^{-1} = \sum_{l=1}^{V} p_l^2$. Define $h = ln\ u$. Then (8) (see also (6)) implies

$$EH_n \le 1 + \frac{1}{h}\, ln\ \frac{n(n-1)}{2} + \frac{1}{u-1}$$

and after simple algebra one finds

$$EH_n \le \frac{2}{h}\, ln\ n + 1 + \frac{1 - ln\ 2}{h} + O(n^{-1}). \tag{10}$$

Hence by (9) $EH_n \sim \frac{2}{h}\, ln\ n = 2\, lg_u\ n$, that is,

$$\lim_{n \to \infty} \frac{EH_n}{ln\ n} = \frac{2}{h}. \tag{11a}$$

How tight is the upper bound (10) ? For binary symmetric tries ($h = ln\ 2$) Devroy proved that [4]

$$EH_n \le 2\, lg_2\, n + 1 + \frac{\gamma - ln\ 2}{ln\ 2} \tag{11b}$$

hence the upper bound (10) is greater than (11) by 0.61. On the other hand, Flajolet [2] shows that for binary symmetric tries

$$EH_n = 2\, lg_2\, n + \frac{\gamma - ln 2}{ln\ 2} + P(\ln n) + o(1) \tag{12}$$

where $P(\ln n)$ is a periodic function with very small amplitude. The derivation of (11) and (12) require, however, much more advanced techniques. In both cases the average EH_n was obtained through the analysis of the asymptotic approximation of the distribution function of H_n.

Some remarks on the asymptotic distribution of H_n

In the subsection we offer some remarks on the asymptotic distribution of H_n. We do not pretend to present rigorous proofs. Rather, we give some reasons justifying the form of the asymptotic distribution.

Assume first that $X_1, X_2, \ldots, X_n$ are identically independently distributed random variables with distribution function $F(x)$. Let also $X_{(n)} = \max\{X_1, \ldots, X_n\}$. It is shown [12], [13] that there exist constants a_n and b_n such that $(X_{(n)} - a_n)/b_n$ has a proper distribution $\Lambda(x)$, as n tends to infinity. In fact, it is proved that the extreme distribution $\Lambda(x)$ may have three different forms. If X_i is exponentially distributed with parameters λ, then $\Lambda(x) = \exp[-e^{-x}]$, that is [12], [13]

$$\lim_{n \to \infty} Pr\{(X_{(n)} - \frac{ln\ n}{\lambda})\ \lambda < x\} = \Lambda(x) = \exp(-e^{-x}]. \tag{13}$$

The situation is a little more delicate for discrete random variables. Anderson [14] showed that if X_i is geometrically distributed with parameter p, then

$$\begin{aligned} \Lambda(x-1) \le \lim_{n \to \infty} \inf\ Pr\{(X_{(n)} - \frac{ln\ n}{ln\ p^{-1}}) ln\ p^{-1} < x\} \le \\ \lim_{n \to \infty} \sup\ Pr\{(X_{(n)} - \frac{ln\ n}{ln\ p^{-1}} ln\ p^{-1} < x\} \le \Lambda(x). \end{aligned} \tag{14}$$

From the practical view point, the difference between (13) and (14) may be ignored if one assumes $\lambda = ln\ p^{-1}$ (i.e., one approximates the geometric distribution with parameter p by the exponential distribution with parameter $\lambda = ln\ p^{-1}$). It is also proved [13] that under some assumptions (13) holds for dependent random variables $X_1, X_2, \ldots, X_n$.

The height of a trie, H_n, is given by (2), where $Y_1, Y_2, \ldots, Y_m$, $m = n(n-1)/2 - n^2$ are dependent random variables geometrically distributed with parameter $u^{-1} = \sum_{l=1}^{V} p_i^2$. Approximating the geometric distribution by the appropriate exponential distribution with parameter $h = ln\ n$ and using (13), one may show that

$$\lim_{n \to \infty} Pr\{H_n < x + 2\ lg_u\ n\} = \exp[-\exp(-x\ ln\ u)]. \tag{15}$$

A rigorous proof of (15) is given in [6], however, quite a different approach is adopted there. The discrete version of (15) for binary symmetric tries can be found in [4].

Note that for large n (15) implies the following approximation

$$Pr\{H_n < x\} \approx \exp\{-\exp[-ln\ u(x - 2\ lg_u\ n)]\}. \tag{16}$$

Let Z be a random variable with the distribution function $\Lambda[(x-\xi)\lambda] = \exp\{-\exp[-(x-\xi)\lambda]\}$. Then, it is shown [15] that $EZ = \xi + \gamma/\lambda$, var $X = \frac{\pi^2}{6\lambda^2}$, where $\gamma = 0.577$ is the Euler constant. By (16) we find that for large n

$$EH_n \approx \frac{2}{h}\ ln\ n + \frac{\gamma - ln 2}{h} + 1 \tag{17a}$$

$$\text{var}\ H_n \approx \frac{\pi^2}{6h^2} \tag{17b}$$

Flajolet in [2] proved that for binary symmetric tries the approximation (17a) is different from the exact asymptotic expression by a fluctuating function with a small amplitude. He also found that the variance, var H_n, is not a constant, but rather a fluctuating function.

3. GENERALIZATIONS

In this section, we generalize the results from Section 2, that is, we investigate Poisson model, consider b-tries ($b > 1$), and finally present some results for dependent keys.

3.1 Poisson model

We replace assumption (iii) by

(iii'). The number of records stored in a trie, N, is a random variable distributed according to Poisson with parameter μ, i.e.,

$$Pr\{N = n\} = \frac{\mu^n}{n!} e^{-\mu}. \tag{18}$$

Under (iii') the Bernoulli model becomes the Poisson model. Let H_μ, H_n denote the height in the Poisson and Bernoulli models, respectively. Then

$$EH_\mu = \sum_{n=0}^{\infty} EH_n \frac{\mu^n}{n!} e^{-\mu}$$

and using (10) we find

$$EH_\mu \le \frac{2}{h} e^{-\mu} \sum_{n=1}^{\infty} ln\ n \frac{\mu^n}{n!} + 1 + \frac{1 - ln2}{h}. \tag{19}$$

To evaluate the series in (19) we use the inequality $ln\ n \le \chi_n$, where χ_n is the n-th Harmonic number. It is known that [16], [17]

$$\sum_{n=1}^{\infty} \chi_n \frac{x^n}{n!} = \int_0^1 \frac{e^x - e^{xy}}{1-y} dy$$

hence

$$e^{-\mu} \sum_{n=1}^{\infty} ln\ n \frac{\mu^n}{n!} \le e^{-\mu} \sum_{n=1}^{\infty} \chi_n \frac{\mu^n}{n!} = \int_0^\mu \frac{1 - e^{-y}}{y} dy = ln\ \mu + \gamma + E_1(\mu) \tag{20}$$

where $E_1(\mu)$ is the exponential integral defined as $E_1(x) = \int_x^\infty e^{-t} t^{-1} dt$ ($|\arg x| < \pi$). Thus, (11) and (20) implies that

$$EH_\mu \le \frac{2}{h} \ln \mu + \frac{E_1(\mu) + \gamma + 1 - \ln 2}{h} + 1 \tag{21}$$

Also, by (11) and (20) we find

$$\lim_{\mu \to \infty} \frac{EH_\mu}{ln\ \mu} = \frac{2}{h}, \tag{22}$$

i.e., $EH_\mu \sim 2\, lg_u\, \mu$.

3.2 The average height of b-tires

We now drop assumption (iv), and consider b-tries with $b > 1$, that is, each external node may store at most b keys. Let $X_1, X_2, \ldots, X_n$ be the keys, and for $i_1, i_2, \ldots, i_{b+1} \in \{1, 2, \ldots, n\}$ we denote $Y(i_1, i_2, \ldots, i_{b+1}) = \mathrm{com}(X_{i_1}, X_{i_2}, \ldots, X_{i_{b+1}})$ the common prefix for $X_{i_1}, \ldots, X_{i_{b+1}}$, i.e., the number of digits that $X_{i_1}, \ldots, X_{i_{b+1}}$ agree. Note that we have $\binom{n}{b+1}$ random variables $Y(i_1, \ldots, i_{b+1})$, and for simplicity we sometimes renumber them and denote $Y_1, Y_2, \ldots, Y_m, m = \binom{n}{b+1}$. Figure 2 shows 2-tries for the same keys as in Figure 1.

A = 000
B = 010
C = 012
D = 100
E = 200
F = 221

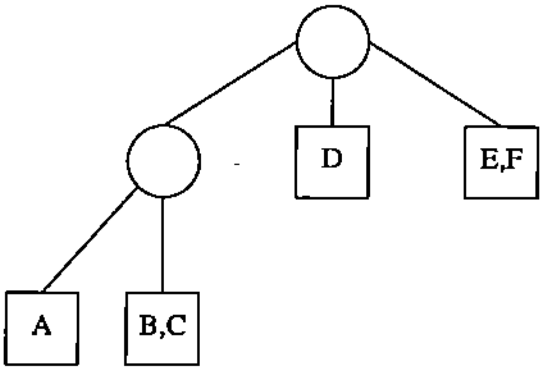

Figure 2. Example of 3-ary digital 2-trie with n=6.

Note that the height of b-tries is given by

$$H_n = 1 + \max_{1 \le i \le m} \{Y_i\} \tag{23}$$

as in (2). The distribution of $Y(i_1, \ldots, i_{b+1})$ is geometric with parameter $u^{-1} = \sum_{l=1}^{V} p_l^{b+1}$, that is

$$Pr\{Y(i_1, \ldots, i_{b+1}) = k\} = u^{-k}(1 - u^{-1}) \quad k = 0, 1, \ldots . \tag{24}$$

Let also $h = ln\ u$.

To compute EH_n we need $E \max_{1 \le i \le m} Y_i$. By (4) and (5) we find

$$E \max_{1 \le i \le m} Y_i \le a_m + m \sum_{k = a_m}^{\infty} u^{-(k+1)} \tag{25}$$

where

$$a_m = \frac{ln\ m}{h} = \frac{ln \left(\begin{array}{c} n \\ b+1 \end{array} \right)}{h}. \tag{26}$$

Note that

$$m \sum_{h = a_m}^{\infty} u^{-(k+1)} \le \frac{1}{h}$$

and

$$\left(\begin{array}{c} n \\ b+1 \end{array} \right) = \frac{n^{b+1}}{(b+1)!} (1 - \frac{1}{n})(1 - \frac{2}{n}) \cdots (1 - \frac{b}{n})$$

hence, these and (25), (26) imply

$$EH_n \le \frac{b+1}{h}\ ln\ n + 1 + \frac{1 - ln(b+1)!}{h} + O(n^{-1}). \tag{27}$$

Since the geometric distribution satisfies (9), we also obtain

$$\lim_{n \to \infty} \frac{EH_n}{ln\ n} = \frac{b+1}{h}. \tag{28}$$

For symmetric V-ary trie, i.e., $p_1 = p_2 = \cdots p_V = \frac{1}{V}$ one immediately shows that $h = b\ ln\ V$

and

$$EH_n \sim (1 + \frac{1}{b})\ lg_V\ n \tag{29}$$

which generalizes Flajolet's result.

To evaluate the asymptotic distribution of H_n we may use the arguments from the preceding section. Then

$$\lim_{n \to \infty} Pr\{H_n < x + (b+1)\ lg_u\ n\} = \exp\{-\exp[-x\ ln\ u]\} \tag{30}$$

which agrees with the result obtained in [6].

3.3 Dependent keys

In many applications the keys are statistically dependent, e.g. see *position trees* and *substring identifiers* (suffix trie) [1]. In this subsection, we relax assumption (ii) keeping the other assumptions unchanged (for simplicity we set $b = 1$).

Let X_k^i, X_l^j denote the i-th and the j-th digit in the k-th and the l-th keys. We assume that there is a dependency between X_k^i, X_l^j, which we express in terms of the joint distribution, that is,

$$p_{n,m}(k,l) \overset{def}{=} Pr\{X_k^i = \alpha_n, \quad X_l^j = \alpha_m\} < 1 \tag{31}$$

where $k,l = 1, 2, \ldots, n,\ n,m = 1, 2, \ldots, V,\ i,j = 1, , \ldots.$ The probability (31) does not depend on i and j because of the assumption (i). Define $Y_{kl} = \mathrm{com}(X_k, X_l)$ as the common prefix of X_k and X_l. By assumption (i) the distribution of Y_{kl} is geometric with parameter $[u(k,l)]^{-1} = \sum_{i=1}^{V} p_{ii}^2(k,l)$. Note that now Y_{kl} are not identically distributed, and $F_{kl}(j) = Pr\{Y_{kl} < j\} = 1 - [u(k,l)]^{-(j+1)}$. The height of the trie is given by (2), and to compute EH_n we need $\max_{k,l} Y_{k,l}$

We now generalize formula (3)–(5). Since inequality (3) holds we also have

$$E \max_{k,l} Y_{kl} \le a_m + \sum_{k,l=1}^{n} \sum_{j=a_n}^{\infty} [1 - F_{kl}(j)] \tag{32}$$

where $m = n(n-1)/2$. The RHS of (32) is minimized for such a_m that

$$\sum_{k=1}^{n} \sum_{l=k+1}^{n} F_{kl}(a_m) = n - 1 \tag{33}$$

For geometric distribution with parameter $u^{-1}(k,l)$ (33) is equivalent to

$$\sum_{k=1}^{n} \sum_{l=k+1}^{n} [u(k,l)]^{-(a_n+1)} = 1 \tag{34}$$

and one finds that

$$a_m \leq \frac{ln\ m}{ln\,[\min\limits_{k,l} n(k,l)]} \tag{35}$$

Let $u_{\min} = \min\limits_{k,l} \{u(k,l)\} = \left[\max\limits_{k,l} \{\sum\limits_{i=1}^{V} p_{\dot{u}}^2(k,l)\} \right]^{-1}$, and $h_{\min} = ln\ u_{\min}$. Noting that $m = n(n-1)/2$ and the contribution of the sum in (32) is O(1) we prove that

$$EH_n \leq \frac{2}{h_{\min}}\ ln\ n + O(1), \tag{36}$$

hence $EH_n = O(ln\ n)$. This result is easy to generalize to b-tries and Poisson model. Note also that the assumption $p_{n,m}(k,l) < 1$ (see (31))is important. For example, if one builds a prefix tree (the k-th key is the prefix of the $(k-1)$st key), then the height is equal to n.

4. CONCLUSIONS

This paper studies the average height of digital tries. Using elementary methods, in contrast to the previous analysis, we proved that $EH_n \sim 2\ lg_u\ n$, where u is a parameter dependent on the distributions of the digits in a key. This result is next extended to Poisson model, b-tries and dependent keys. Under quite general assumptions, we show that $EH_n = O(ln\ n)$, even for dependent keys. The methodology proposed here can be applied to analyze some other quantities of the digital tries, as well as to obtain some characteristics for Patricia tries and digital trees.

REFERENCES

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

[2] Flajolet, Ph., *On the performance evaluation of extendible hashing and trie searching*, Acta Informatica, 20, (1983), pp. 345–369.

[3] Jacquet, Ph. and Regnier, M., *Trie partitioning process: Limiting distributions*, Proc. of CAAP'86.

[4] Devroye, L., *A probabilistic analysis of the height of tries and of the complexity of trie sort*, Acta Informatica, 21, (1984), pp. 229–232.

[5] Pittel, B., *Asymptotic growth of a class of random trees*, The Annalus of Probability, 13, (1985), pp. 414–427.

[6] Pittel, B., *Path in a random digital tree: Limiting distributions*, Adv. Appl. Probl., 18, (1986), pp. 139–155.

[7] Regnier, M., *On the average height of trees in digital searching and dynamic hashing*, Inform. Processing Lett., 13, (1981), pp. 64–66.

[8] Yao, A., *A note on the analysis of extendible hashing*, Inform. Processing Lett., 11, (1980), pp. 84–86.

[9] Szpankowski, W., *Average complexity of additive properties for multiway tries: A unified approach*, Proc. CAAP'87, (1987).

[10] Lai, T. and Robbins, H., *Maximally dependent random variables*, Proc. Nat. Acad. Sci. USA, 73, (1986), pp. 286–288.

[11] Lai, T. and Robbins, H., *A class of dependent random variables and their maxima*, Z. Wahrscheinlichkeitscheorie, 42, (1978), pp. 89–111.

[12] David, H., *Order statistics*, John Wiley & Sons, New York, (1980).

[13] Galambos, J., *The asymptotic theory of extreme order statistics*, John Wiley & Sons, New York, (1978).

[14] Anderson, C., *Extreme value theory for a class of discrete distributions with applications to some stochastic processes*, J. Appl. Prob., 7, (1970), pp. 59–113.

[15] Johnson, M. and Kotz, S., *Continuous univariate distributions*, Houghton Mefflin Company, New York, (1970).

[16] Knuth, D., *The art of computer programming sorting and searching*, Addison-Wesley, (1973).

[17] Szpankowski, W., *On an asymptotic analysis of a tree-type algorithm for broadcast communications*, Inform. Processing Lett., (1986).

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

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