1989

A Note on the Height of Suffix Trees

Luc Devroye

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

Bonita Rais

Report Number:
89-905

A NOTE ON THE HEIGHT OF SUFFIX TREES

Luc Devroye
Wojciech Szpankowski
Bonita Rais

CSD-TR-905
September 1989

A NOTE ON THE HEIGHT OF SUFFIX TREES

Luc Devroye

School of Computer Science
McGill University

and

Wojciech Szpankowski and Bonita Rais

Department of Computer Science
Purdue University

Abstract

Consider a word in which the individual symbols are independent integers occurring with probabilities p_i, and let H_n be the height of the suffix tree constructed from the first n suffixes of this word. We show that H_n is asymptotically close to $2 \log n / \log(1/\sum_i p_i^2)$ in many respects: the difference is $O(\log \log n)$ in probability, and the ratio tends to one in probability and in the mean.

Keywords and phrases: Keywords and phrases Suffix tree, height, trie hashing, analysis of algorithms, weak convergence.

CR Categories: 3.74, 5.25, 5.5.

Acknowledgement : Research of the first two authors was carried out at INRIA, Rocquencourt, France. Research of the first author was sponsored by NSERC Grant A3456 and by FCAC Grant EQ-1678. Research of the the second author was supported by NATO Collaborative grant 0057/89, and NSF grants NCR-8702115 and CCR-8900305. The last author was partially supported by NSF grant CCR-8900305.

Addresses : Luc Devroye, School of Computer Science, McGill University, 3840 University Street, Montreal, Canada H3A 2A7. Wojciech Szpankowski and Bonita Rais, Department of Computer Science, Purdue University, West Lafayette, IN. 47907, USA.

1. Introduction.

Tries are efficient data structures that were developed and modified by Fredkin (1960), Knuth (1973), Larson (1978), Fagin, Nievergelt, Pippenger and Strong (1979), Litwin (1981, 1985), Aho, Hopcroft and Ullman (1983) and others. Multidimensional generalizations were given in Nievergelt, Hinterberger and Sevcik (1984) and Régnier (1985). One kind of trie, the suffix tree, is of particular utility in a variety of algorithms on strings (Aho, Hopcroft and Ullman (1975), McCreight (1976), Apostolico (1985)). However, except for the results in Apostolico and Szpankowski (1987), who give an upper bound on the expected height (see also Szpankowski (1988)), very little is known about the expected behavior of suffix trees. Also noteworthy is a result by Blumer, Ehrenfeucht and Haussler (1989) who obtained asymptotics for the expected size of the suffix tree under an equal probability model. The difficulty arises from the interdependence between the keys, which are suffixes of one string. In this note, we study the height of the suffix tree. The results of our analysis find applications in many areas (Aho, Hopcroft and Ullman (1975), Apostolico (1985)). For example, suffix trees are used as a unifying framework for a class of linear time sequential data compression, and they are employed to decide whether a word contains a square subword, to spot all such squares, to allocate statistics without overlap of all subwords in a textstring, and so forth. Consequences of our findings for an efficient design of algorithms are extensively discussed in Apostolico and Szpankowski (1988).

We consider an i.i.d. sequence $X_1, X_2, \ldots$ of integer-valued nonnegative random variables with $\mathbf{P}(X_1 = i) = p_i$ for $i = 0,1,2,\ldots$ and $\sum_i p_i = 1$. The X_i's should be considered as symbols in some alphabet. We do not assume that the alphabet is finite, but we will assume that no p_i is one, for otherwise all the symbols are identical with probability one. The **suffixes** Y_i are obtained by forming the sequences $Y_i = \{X_i, X_{i+1}, \ldots\}$. The **suffix tree** based upon $Y_1, \ldots, Y_n$ is nothing but the trie obtained when the Y_i's are used as words (for a definition of tries, see Knuth (1973); for a survey of recent results, see Szpankowski (1988)). Note however that we do not compress the trie as in a PATRICIA trie, i.e. no substrings are collapsed into one node.

In this note we study the **height** H_n of the suffix tree, which is nothing but

$$H_n = \max_{i \neq j, 1 \leq i, j \leq n} C_{ij} ,$$

where C_{ij} is the length of the common prefix of Y_i and Y_j, i.e. $C_{ij} = k$ if $(X_i, \ldots, X_{i+k-1}) = (X_j, \ldots, X_{j+k-1})$ and $X_{i+k} \neq X_{j+k}$. In the announcement and derivation of the results, we will need the metrics $||p||_r = (\sum_i p_i^r)^{1/r}$, $0 < r < \infty$ and $||p||_\infty = \max_i p_i$. Finally, to alleviate the notation, we define $Q = 1/||p||_2$.

Theorem.

For the suffix tree, $H_n / \log n \to 1/\log Q$ in probability. Also, for all $m \geq 1$, $\mathbf{E}H_n^m \sim \log^m n / \log^m Q$.

We will prove this result using only elementary probability theoretical tools, such as the second moment method. Nevertheless, we will in fact be able to show that

$$\mathbf{P}(|H_n - \frac{\log n}{\log Q}| > (1+\varepsilon)\log\log n) \to 0$$

for all $\varepsilon > 0$. Thus, the variations of H_n are at best of the order of $\log\log n$.

It is interesting to note that the first asymptotic term ($\log n/\log Q$) is of the same order of magnitude as for the asymmetric trie obtained if the words $Y_1, \ldots, Y_n$ had been i.i.d. (Pittel, 1985, 1986; Szpankowski, 1988). In 1985, Pittel showed that $H_n/\log n \to 1/\log Q$ almost surely, and in 1986, he showed that $H_n - \log n/\log Q = O(1)$ in probability. Other properties of the height of a trie under the independent model can be found in Yao (1980), Régnier (1981), Flajolet (1983), Devroye (1984), Pittel (1985,1986), Jacquet and Régnier (1986), and Szpankowski (1988), who presents a survey of recent results. The reader is also referred to some other related papers such as Kirschenhofer and Prodinger (1986), Flajolet and Puech (1986), Flajolet and Sedgewick (1986) and Szpankowski (1988b).

2. Preliminary results.

We present four simple lemmata. The first two are trivial. The third one is due to Apostolico and Szpankowski (1987).

Lemma 1.

$$||p||_\infty^2 \le ||p||_2^2 \le ||p||_\infty .$$

Lemma 2. For every $r \ge 2$ the following holds

$$||p||_r \le ||p||_2$$

Proof of Lemma 2.

Let us consider a function $f(x) = \{\sum_i p_i^x\}^{1/x}$ for $x > 0$. Then, it is easy to show that the first derivative of $f(x)$ is negative for all $x > 0$. This completes the proof. For more details see Szpankowski (1988), and Karlin and Ost (1985). ■

Lemma 3.

For $0 < |i-j| = d < k$, we have

$$\mathbf{P}(C_{ij} \ge k) = (\sum_i p_i^{l+2})^r \, (\sum_i p_i^{l+1})^{d-r} ,$$

where $l = [k/d]$ ([.] denotes the integer fraction), and $r = k - dl = k \bmod d$. Also, for $|i-j| \ge k$, we have $\mathbf{P}(C_{ij} \ge k) = ||p||_2^{2k}$.

Lemma 4.

For $0 < |i-j| = d < k$, we have $\mathbf{P}(C_{ij} \ge k) \le ||p||_2^{k+d}$.

Proof of Lemma 4.

In the notation of Lemma 3, and using Lemma 2 we immediately obtain

$$\mathbf{P}(C_{ij} \geq k) = (\sum_i p_i^{l+2})^r \, (\sum_i p_i^{l+1})^{d-r} \leq ||p||_2^{(l+2)r+(l+1)(d-r)} = ||p||_2^{k+d} . \blacksquare$$

3. Proof of the Theorem.

We prove our theorem by showing two tight bounds for the hight H_n. Roughly speaking, we shall show that for every ε and large n the following holds: $\mathbf{P}(H_n > (1+\varepsilon)\cdot\log_Q n) \to 0$ as $n \to \infty$ (upper bound), and $\mathbf{P}(H_n < (1-\varepsilon)\cdot\log_Q n) \to 1$ as $n \to \infty$ (lower bound).

We start with an easier part of our proof, namely the upper bound. Assume that $2 \leq k \leq n-1$. We have from Lemmata 2 and 4, and Bonferroni's inequality

$$\begin{aligned}\mathbf{P}(\max_{i \neq j} C_{ij} \geq k) &\leq 2n(\sum_{d=1}^{k-1} \mathbf{P}(C_{1,1+d} \geq k) + \sum_{d=k}^{n-1} \mathbf{P}(C_{1,1+d} \geq k)\,) \\ &\leq 2n(\sum_{d=1}^{k-1} ||p||_2^{k+d} + \sum_{d=k}^{n-1} ||p||_2^{2k}) \\ &\leq 2n(\frac{||p||_2^{k+1}}{1-||p||_2} + n||p||_2^{2k})\,. \qquad (1)\end{aligned}$$

This tends to zero provided that $||p||_2 < 1$ (this is always true) and that $n||p||_2^k \to 0$ (for this, it suffices that $k = (\log n + \omega_n)/(-\log||p||_2)$, with $\omega_n \to \infty$). This establishes the weak upper bound of the Theorem. Note also that, by (1),

$$\begin{aligned}&\mathbf{E}(\max_{i \neq j} C_{ij} \log(\frac{1}{||p||_2}) - \log n)_+^m \\ &= \int_0^\infty \mathbf{P}(\max_{i \neq j} C_{ij} \log(\frac{1}{||p||_2}) - \log n > u^{1/m})\, du \\ &\leq \int_0^\infty (\frac{2e^{-u^{1/m}}}{1-||p||_2} + \frac{2e^{-2u^{1/m}}}{||p||_2}\,)\, du < \infty\,.\end{aligned}$$

Thus, $\mathbf{E}H_n \leq (\log n + A(m,p))/\log Q$ for some finite constant $A(m,p)$.

A matching lower bound is obtained by the second moment method. We will use a form due to Chung and Erdös (1952), which states that for events A_i, we have

$$\mathbf{P}(\bigcup_i A_i) \geq \frac{(\sum_i \mathbf{P}(A_i))^2}{\sum_i \mathbf{P}(A_i) + \sum_{i \neq j} \mathbf{P}(A_i \cap A_j)}\,.$$

Let S be the collection of pairs of indices (i,j) with $1 \leq i,j \leq n$, and $|i-j| \geq k$. Let

$A_{ij} = [C_{ij} \ge k]$. Then

$$\mathbf{P}(\max_{i \ne j} C_{ij} \ge k) \ge \mathbf{P}(\bigcup_{(i,j)\in S} A_{ij})$$

$$\ge \frac{\left(\sum_{(i,j)\in S} \mathbf{P}(A_{ij})\right)^2}{\sum_{(i,j)\in S} \mathbf{P}(A_{ij}) + \sum_{(i,j),(l,m)\in S;(i,j)\ne(l,m)} \mathbf{P}(A_{ij} \cap A_{lm})} .$$

To prove our lower bound it is enough to show that the probability on the RHS of the above tends to 1 for k "slightly" larger than $\log_Q n$ ($k = \log_Q n + \omega_n$). First we note that when $k = o(n)$, then

$$\sum_{(i,j)\in S} \mathbf{P}(A_{ij}) = |S| \; ||p||_2^{2k} \sim n^2 ||p||_2^{2k} \text{ (Lemma 1) .}$$

We decompose the collection of pairs of pairs of indices $\{((i,j),(l,m)):(i,j)\in S, (l,m)\in S, (i,j)\ne(l,m)\}$ as follows into $I_1 \cup I_2 \cup I_3$: I_1 captures all members with $\min(|l-i|,|l-j|) \ge k$ and $\min(|m-i|,|m-j|) \ge k$. I_2 holds all members with either $\min(|l-i|,|l-j|) \ge k$ and $\min(|m-i|,|m-j|) < k$, or $\min(|l-i|,|l-j|) < k$ and $\min(|m-i|,|m-j|) \ge k$. Finally, I_3 collects all members with $\min(|l-i|,|l-j|) < k$ and $\min(|m-i|,|m-j|) < k$. By Lemmata 1 and 2,

$$\sum_{((i,j),(l,m))\in I_1} \mathbf{P}(A_{ij} \cap A_{lm}) \le n^4 ||p||_2^{4k} ,$$

$$\sum_{((i,j),(l,m))\in I_2} \mathbf{P}(A_{ij} \cap A_{lm}) \le 8k\, n^3 ||p||_2^{2k} \; ||p||_\infty^{k} \le 8k\, n^3 ||p||_2^{3k} ,$$

$$\sum_{((i,j),(l,m))\in I_3} \mathbf{P}(A_{ij} \cap A_{lm}) \le (4k)^2 \, n^2 ||p||_2^{2k} .$$

If we choose k such that $n\, ||p||_2^{k}/k \to \infty$, then indeed

$$\sum_{(i,j),(l,m)\in S;(i,j)\ne(l,m)} \mathbf{P}(A_{ij} \cap A_{lm}) \sim n^4 ||p||_2^{4k} .$$

Collecting all this shows that $\mathbf{P}(H_n \ge k) \to 1$ when $n \to \infty$. Note that we can take $k = [(\log n - (1+\varepsilon)\log\log n)/(-\log ||p||_2)]$ for $\varepsilon > 0$. Also,

$$\mathbf{E}H_n \ge k\; \mathbf{P}(H_n \ge k) - k$$

if k is chosen as indicated. This concludes the proof of the lower bound and of the Theorem. ■

4. References.

A.V. Aho, J.E. Hopcroft, and J.D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass., 1975.

A.V. Aho, J.E. Hopcroft, and J.D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, Mass., 1983.

A. Apostolico, "The myriad virtues of suffix trees," in *Combinatorial Algorithms on Words*, pp. 85-96, Springer-Verlag, 1985.

A. Apostolico and W. Szpankowski, "Self-alignments in words and their applications," Technical Report CSD-TR-732, Department of Computer Science, Purdue University, 1987.

A. Blumer, A. Ehrenfeucht, and D. Haussler, "Average sizes of suffix trees and DAWGs," Technical Report, 1989.

K.L. Chung and P. Erdös, "On the application of the Borel-Cantelli lemma," *Transactions of the American Mathematical Society*, vol. 72, pp. 179-186, 1952.

L. Devroye, "A probabilistic analysis of the height of tries and of the complexity of triesort," *Acta Informatica*, vol. 21, pp. 229-237, 1984.

R. Fagin, J. Nievergelt, N. Pippenger, and H.R. Strong, "Extendible hashing - a fast access method for dynamic files," *ACM Transactions on Database Systems*, vol. 4, pp. 315-344, 1979.

P. Flajolet, "On the performance evaluation of extendible hashing and trie search," *Acta Informatica*, vol. 20, pp. 345-369, 1983.

P. Flajolet and C. Puech, "Tree structure for partial match retrieval," *Journal of the ACM*, vol. 33, pp. 371-407, 1986.

P. Flajolet and R. Sedgewick, "Digital search trees revisited," *Siam Journal on Computing*, vol. 15, pp. 748-767, 1986.

E.H. Fredkin, "Trie memory," *Communications of the ACM*, vol. 3, pp. 490-500, 1960.

P. Jacquet and M. Régnier, "Trie partitioning process: limiting distributions," in *Lecture Notes in Computer Science*, vol. 214, pp. 196-210, 1986.

S. Karlin and F. Ost, "Some monotonicity properties of Schur powers of matrices and related inequalities," *Linear Algebra and Applications*, vol. 68, pp. 47-65, 1985.

P. Kirschenhofer and H. Prodinger, "Some further results on digital trees," in *Lecture Notes in Computer Science*, vol. 226, pp. 177-185, Springer-Verlag, Berlin, 1986.

D.E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.

P.A. Larson, "Dynamic hashing," *BIT*, vol. 18, pp. 184-201, 1978.

W. Litwin, "Trie hashing," in *Proceedings of the ACM-SIGMOD Conference on MOD*, Ann Arbor, MI, 1981.

W. Litwin, "Trie hashing: further properties and performances," in *Proceedings of the International Conference on Foundations of Data Organization*, Kyoto, 1985.

E.M. McCreight, "A space-economical suffix tree construction algorithm," *Journal of the ACM*, vol. 23, pp. 262-272, 1976.

J. Nievergelt, H. Hinterberger, and K.C. Sevcik, "The grid file : an adaptable, symmetric multikey file structure," *ACM Transactions on Database Systems*, vol. 9, pp. 38-71, 1984.

B. Pittel, "Asymptotical growth of a class of random trees," *Annals of Probability*, vol. 13, pp. 414-427, 1985.

B. Pittel, ''Path in a random digital tree: limiting distributions,'' *Advances in Applied Probability*, vol. 18, pp. 139-155, 1986.

M. Régnier, ''On the average height of trees in digital searching and dynamic hashing,'' *Information Processing Letters*, vol. 13, pp. 64-66, 1981.

W. Szpankowski, ''On the height of digital trees and related problems,'' Technical Report CSD-TR-816, Department of Computer Science, Purdue University, 1988.

W. Szpankowski, ''Some results on *V*-ary asymmetric tries,'' *Journal of Algorithms*, vol. 9, pp. 224-244, 1988.

A. Yao, ''A note on the analysis of extendible hashing,'' *Information Processing Letters*, vol. 11, pp. 84-86, 1980.