Compact suffix trees resemble PATRICIA tries: limiting distribution of depth

Philippe Jacquet, B. Rais, Wojciec Szpankowski

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Compact Suffix Trees Resemble Patricia Tries: Limiting Distribution of Depth

Philippe JACQUET
Bonita RAIS
Wojciech SZPANKOWSKI

N° 1995
Juillet 1993

PROGRAMME 2

Calcul symbolique,
programmation
et génie logiciel

COMPACT SUFFIX TREES RESEMBLE PATRICIA TRIES: LIMITING DISTRIBUTION OF DEPTH

Revised: June 26, 1993

Philippe Jacquet*
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France

Bonita Rais†
Dept. of Computer Science
Ball State University
Muncie, IN 47306
U.S.A.

Wojciech Szpankowski‡
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

Suffix trees are the most frequently used data structure in algorithms on words. Despite this, little is known about their behavior in a probabilistic framework. In this paper, we consider the depth of a compact suffix tree, also known as the PAT tree, under some simple probabilistic assumptions. In fact, for the case of an asymmetric alphabet, we prove that the limiting distribution for the depth in a PAT tree is the same as the limiting distribution for the depth in a PATRICIA trie, even though the PATRICIA trie is constructed over statistically independent strings. In other words, the limiting distribution for the depth in a PAT tree storing n suffixes is normal.

LES ARBRES SUFFIXES COMPACTES RESSEMBLENT AUX ARBRES PATRICIA: DISTRIBUTION LIMITE DES PROFONDEURS

Résumé

Les arbres suffixes sont les structures de données les plus fréquemment employés pour les algorithmes sur les mots. En dépis de cela, on connaît peu de choses sur leur comportement sous des modèles probabilistes. Dans ce papier, nous considérons les profondeurs dans un arbre suffixe compacté, structure aussi comm sous le nom d'arbre PAT, et ceci sous un modèle probabiliste simple. Nous montrons que la distribution limite des profondeurs dans l'arbre PAT rejoint la distribution limite des profondeurs dans l'arbre PATRICIA, bien que ce dernier soit construit sur des mots indépendants. En conséquence, nous en déduisons que les profondeurs dans l'arbre PAT sont asymptotiquement distribuées selon la loi normale.

*This research was primary supported by NATO Collaborative Grant 0057/89.

†This research was in part supported by AFOSR grant 90-0107 and by NSF grant CCR-8900305.

‡This author's research was supported in part by NATO Collaborative grant 00570/89, and in part by AFOSR grant 90-0107, by NSF grants CCR-9201078 and NCR-9206315, and by grant R01 LM05118 from the National Library of Medicine. This paper was revised while the author was visiting INRIA, Rocquencourt, France, and he wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support.

1. INTRODUCTION

Suffix trees have found a wide variety of applications in algorithms on words including: the longest repeated substring [16], squares or repetitions in strings [1], string statistics [1], string matching [4], approximate string matching [4], string comparison, compression schemes [9], implementation of Lempel-Ziv algorithm, genetic sequences, biologically significant motif patterns in DNA [4], sequence assembly [4], approximate-overlaps [4], and so forth. It is fair to say that suffix trees are most widely used data structure in algorithms on words. Despite this, very little is known about their behavior in a probabilistic framework. A clear example illustrating the benefits from a probabilistic analysis is given in Chang and Lawler [4], who recently used some elementary property of a typical behavior of suffix trees to design a superfast algorithm for the approximate string matching problem.

In recent years, a resurgence of interest in suffix trees has led to a better understanding of their behavior under probabilistic models. However, most of the probabilistic results concern noncompact suffix trees constructed over a string whose symbols occur independently of each other and/or deal with convergence in probability or almost sure (a.s) convergence. The probabilistic analysis of noncompact suffix trees was initiated by Apostolico and Szpankowski [2] who gave an upper bound for the expected height. The asymptotic height, which provides an improved upper bound, is computed in Devroye, Szpankowski and Rais [5]. The limiting distribution for the depth in a noncompact suffix tree was recently computed by Jacquet and Szpankowski [8]. In [15], Szpankowski obtained some results involving (a.s) convergence for the depth, height, and other related quantities of suffix trees and compact suffix trees for a more general probabilistic model. Also, the external path length of the noncompact suffix tree was analyzed by Shields [13]. The average size of suffix tree was established in [8] (cf. [3]). Guibas and Odlyzko [7] have obtained results concerning the overlapping and periodicity in strings. Finally a survey of results for digital trees is given in a book by Gonnet and Baeza-Yates [6]. It is important to note that previously there were very few known results for the compact suffix tree (cf. [15]). In this paper, we compute the limiting distribution for the depth in a compact suffix tree, providing a characterization of the depth.

Here we give a brief definition of a compact suffix tree, also known as a PAT tree. We begin with a string $X = x_1x_2x_3\ldots$ where x_i is a symbol from the finite alphabet $\Sigma = \{\omega_1, \omega_2, \ldots, \omega_V\}$. In this research, we assume an independent, asymmetric alphabet; in other words, $\Pr\{x_j = \omega_i\} = p_i$ for any j, $\sum_{i=1}^{V} p_i = 1$, and there is at least one i such that $p_i \neq 1/V$. Such a probabilistic model is known as an asymmetric Bernoulli model. The i-th suffix of X is the string given by $X_i = x_ix_{i+1}x_{i+2}\ldots$. In a suffix tree, each suffix is

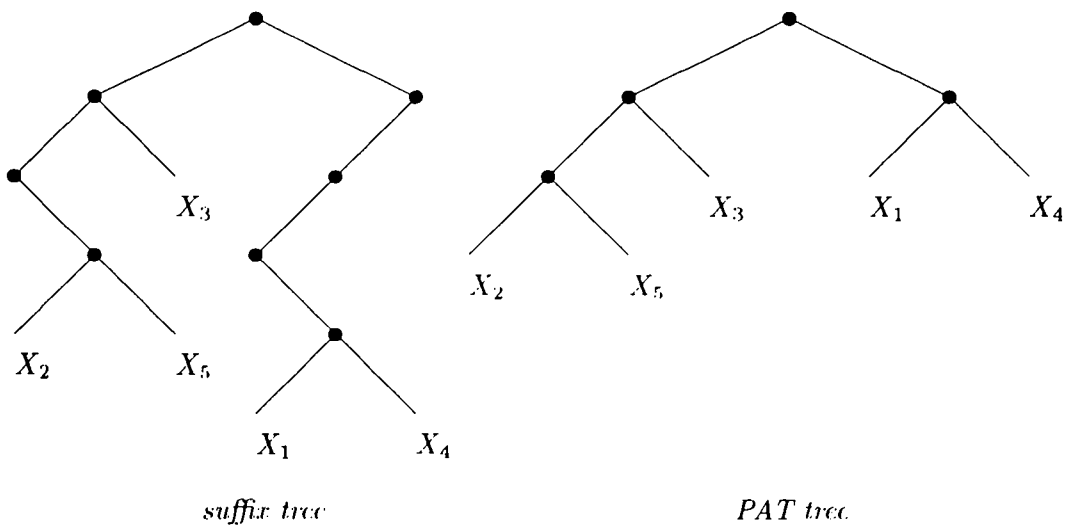

Figure 1: Suffix tree and PAT tree of $X = 10010011\ldots$ for $n = 5$.

stored in a leaf of the tree. The tree is built recursively, splitting into subtrees at the k-th step as determined by the k-th symbol of each suffix. An example of a suffix tree for the string $X = 10010011\ldots$ appears in Figure 1. The PAT tree, as its name implies, is similar to the PATRICIA trie in that all consecutive, non-branching nodes of the suffix tree are collapsed into single node. The corresponding PAT tree also appears in Figure 1.

2. MAIN RESULTS

In this section we present the statement of our main results and its implications. Our results hold under the model in which the string X is an infinite string of symbols from an independent, asymmetric alphabet of V symbols. Let D_n^{PAT} be the depth of the PAT tree constructed over the first n suffixes of X. The depth of any tree is defined to be the depth of a randomly chosen key stored in the tree. Thus,

$$\Pr\{D_n^{PAT} \geq k\} = \frac{1}{n}\sum_{i=1}^{n} \Pr\{D_n^{PAT}(X_i) \geq k\} \tag{1}$$

where $D_n^{PAT}(X_i)$ is the depth of the suffix X_i in a PAT tree with n suffixes. We now state our main result.

THEOREM. *Consider the PAT tree constructed over the first n suffixes of a string X generated over a finite alphabet in the asymmetric Bernoulli model. Then,*

(i) *For large n the average ED_n^{PAT} depth of a PAT tree is*

$$ED_n^{PAT} = \frac{1}{H}\{\log n + \gamma + \frac{H_2}{2H}\} + P_1(\log n) + O(\frac{1}{n^\epsilon})$$

and the variance $varD_n^{PAT}$ of the depth is

$$varD_n^{PAT} = \frac{H^2 - H_2}{H^3}\log n + A + P_2(\log n) + O(\frac{1}{n^\epsilon})$$

where $H = -\sum_{i=1}^{V} p_i \log p_i$ is the entropy of the alphabet, $H_2 = \sum_{i=1}^{V} p_i \log^2 p_i$, $\gamma = 0.577$ is Euler's constant, $P_1(x)$ and $P_2(x)$ are fluctuating, periodic functions of small amplitudes, and A is an explicit constant found in [14].

(ii) *The random variable $\left(\frac{D_n^{PAT} - ED_n^{PAT}}{\sqrt{varD_n^{PAT}}}\right)$ is asymptotically normal with mean zero and variance one, that is,*

$$\lim_{n\to\infty} \Pr\{D_n^{PAT} \le ED_n^{PAT} + x\sqrt{varD_n^{PAT}}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt.$$

Remarks and Observations

(i) *Comparison of the depth in PATRICIA tries and PAT trees.* In this case it appears that the similarities of the trie and the suffix tree carries through into the compact versions of each tree. That is, the PATRICIA trie and the PAT tree have a similar limiting distribution. Again this is somewhat remarkable considering the nature of the data being used. The high dependency among suffixes does not alter the typical shape of the tree too much when compared to a PATRICIA trie. Because of this, we can argue, in much the same way as in [12] for the PATRICIA trie, that the PAT tree is, with high probability, well-balanced.

(ii) *Symmetric case.* Unfortunately we are unable to extend our results for the depth in a PATRICIA trie to the PAT tree in the symmetric model. For the trie, Pittel [10] proved that

$$\lim_{n\to\infty} \sup_x |\Pr\{D_n \le x\} - e^{-nV^{-x}}| = 0$$

uniformly in x, where D_n is the depth in a trie. This same result is obtained by Jacquet and Szpankowski in [8] for the symmetric model of suffix trees. Although the proof as described in [10] for the trie is quite simple, the proof for the PATRICIA tree in the symmetric model is quite complicated, as shown in [12], and at this time, we do not know how to extend it to the PAT tree.

3. ANALYSIS

In analyzing the depth of the PAT tree, we will make use of the result obtained by Rais, Jacquet and Szpankowski in [12] for the depth in a PATRICIA trie, and the result of Jacquet and Szpankowski [8] regarding the limiting distribution for the depth in a suffix tree.

The proof of our theorem will be completed in the steps listed below:

(i) First we will show that $D_n^{PAT} \leq_{st} D_n^S$ stochastically; that is, for any x, we have $\Pr\{D_n^{PAT} \geq x\} \leq \Pr\{D_n^S \geq x\}$, where D_n^S is the depth of a noncompact suffix tree with n keys. This will provide an upper bound for D_n^{PAT} since the limiting distribution of D_n^S is given in [8] with mean ED_n^S and $varD_n^S$ as given in our theorem.

(ii) Second, we will construct a compact tree over a particular subset of size m of suffixes of the given string X. Then, defining the depth of this special tree as D_m^{PAT}, we show that $D_m^{PAT} \leq_{st} D_n^{PAT}$ stochastically. This provides a lower bound.

(iii) Third, we show that D_m^{PAT} and the depth of a PATRICIA trie over m independent keys D_m^P converge to the same distribution. In other words, there exists $\epsilon_m > 0$, such that for all k,

$$|\Pr\{D_m^{PAT} > k\} - \Pr\{D_m^P > k\}| < \epsilon_m .$$

(iv) Finally, we show for our choice of m that D_m^P and D_n^P, the depth of PATRICIA tries with m and n independent keys, respectively, converge to the same distribution. In [12] we have that D_n^P is asymptotically normally distributed with mean ED_n^P and $varD_n^P$ as given in our theorem.

When we have completed these steps, D_n^{PAT} will be bounded by D_n^S and D_n^P which have equivalent limiting distributions. This will show that the limiting distribution of D_n^{PAT} is equal to each of them, and will prove that D_n^{PAT} is normally distributed.

The first step is easy. Clearly, $D_n^{PAT} \leq_{st} D_n^S$ since the depth of any key in a compact suffix tree is at most equal to the depth of that same key in the corresponding suffix tree and, in fact, may be less.

Next, we construct a compact "suffix" tree over a particular set of m suffixes. The m suffixes are chosen in much the same way as in [11, 15] for the computation of the lower bound for the height of a suffix tree. Let $M = \lfloor 2C \log n \rfloor$ where $C \log n$ is the leading term in the asymptotic height of the suffix tree computed in [5, 11, 15] (in fact, $C = -1/\log(p_1^2 + \ldots + p_V^2)$ in the Bernoulli model). Then, we choose $Y_i = X_{M(i-1)+1}$ for

$i = 1, \ldots, m$ where $m = \lceil n/M \rceil = O(\frac{n}{\log n})$. By choosing the Y_i's in this way, they do not overlap one another for the first M symbols, and thus, they are nearly independent. This will make computing the distribution of the depth in this tree much easier than in the PAT tree containing all n suffixes. (Intuitively, the tree can be considered to be a PATRICIA trie rather than a PAT tree, but this will be rigorously proved shortly.) We now prove that $D_m^{PAT} \leq_{st} D_n^{PAT}$ where D_m^{PAT} is the depth of the new tree built over $Y_1, Y_2, \ldots, Y_m$..

Unfortunately, it is not necessarily true that the depth of a tree increases when an additional suffix is added to the tree. This is caused by the fact that the depth of a tree is defined to be the depth of a randomly chosen key as illustrated in (1). However, we can say that $D_m^{PAT}(Y_i) \leq_{st} D_n^{PAT}(Y_i)$ for $i = 1, \ldots, m$ since each Y_i in the tree with m keys is also in the tree with n keys at a depth at least as great as in the tree with m keys. But this also says that $\Pr\{D_m^{PAT}(Y_i) \geq k\} \leq \Pr\{D_n^{PAT}(Y_i) \geq k\}$, which leads to the following sequence of steps:

$$\begin{aligned}
\Pr\{D_n^{PAT} \geq k\} &= \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{m} \Pr\{D_n^{PAT}(X_{M(i-1)+j}) \geq k\} \\
&\geq \frac{m}{n} \sum_{j=1}^{M} \frac{1}{m} \sum_{i=1}^{m} \Pr\{D_m^{PAT}(Y_i) \geq k\} \\
&= \frac{m}{n} \sum_{j=1}^{M} \Pr\{D_m^{PAT} \geq k\} \\
&\geq \Pr\{D_m^{PAT} \geq k\}.
\end{aligned}$$

Thus, D_m^{PAT} is a lower bound for D_n^{PAT}.

We know present a proof that our PAT tree on the specially chosen m suffixes of X is comparable to a PATRICIA trie on m independent keys. To do this, we construct a second tree whose m keys, Y_i^P for $i = 1, \ldots, m$, are given as follows. The key Y_i^P agrees with the key Y_i on the first M symbols and the remaining symbols are chosen arbitrarily. Obviously, this new tree is a PATRICIA trie since the keys are independent. Thus the limiting distribution D_m^P for the depth of this PATRICIA tree with m independent keys is normal and is given in [12].

Finally, by our choice of M, we know that the $\Pr\{h_n > M\} \to 0$ as $n \to \infty$, where h_n is the height of a suffix tree on n keys. This implies that our compact "suffix" tree on m keys and the PATRICIA tree constructed above are identical with probability tending to 1. Thus, the limiting distributions D_m^P and D_m^{PAT} are the same.

Our proof is not yet complete because we cannot equate the limiting distribution of D_m^P with D_n^S. The problem is that, although D_m^P and D_n^S are both normal, D_m^P has mean and

variance of $O(\log m)$ and D_n^S has mean and variance of $O(\log n)$. However, when $k \to \infty$, D_k^P converges to the normal distribution with mean equivalent to $c_1 \log k$ and variance equivalent to $c_2 \log k$. Since $m = \lceil n/(2C \log n) \rceil$ the mean $c_1 \log m = c_1 \log n + o(\sqrt{\log n})$ and the variance $c_2 \log m$ is equivalent to $c_2 \log n$. These facts together with the normal convergence easily lead to the convergence in distribution of D_m^P and D_n^P.

Putting all the above steps together, we have for large n,

$$D_n^P \stackrel{d}{=} D_m^P \stackrel{d}{=} D_m^{PAT} \leq_{st} D_n^{PAT} \leq_{st} D_n^S,$$

where $\stackrel{d}{=}$ denotes equality in distribution. But since D_n^P and D_n^S have the same limiting distribution, D_n^{PAT} also has the same limiting distribution which is given explicitly in our theorem. Our proof is now complete.

References

[1] A. Apostolico. The Myriad Virtue of Suffix Trees. *Springer NATO ASI Ser. F12*, 85–96, March 1985.

[2] A. Apostolico and W. Szpankowski. Self-Alignments in Words and Their Applications. *Journal of Algorithms*, 13, 1992.

[3] A. Blumer, A. Ehrenfeucht, and D. Haussler. Average Sizes of Suffix Trees and DAWGs. *Discrete Applied Mathematics*, 24:37–45, 1989.

[4] W. Chang and E. Lawler. Approximate String Matching in Sublinear Expected Time. *Proceedings of 1990 FOCS*, 116–124, 1990.

[5] L. Devroye, W. Szpankowski, and B. Rais. A Note on the Height of Suffix Trees. *SIAM Journal on Computing*, 21(1):48–53, 1991.

[6] G. H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures*. Addison-Wesley, 1991.

[7] L. Guibas and A. W. Odlyzko. String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory*, Series A(30):183–208, 1981.

[8] P. Jacquet and W. Szpankowski. Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory. Ser. A*, to appear.

[9] A. Lempel and J. Ziv. On the Complexity of Finite Sequences. *IEEE Information Theory*, 22:75–81, 1976.

[10] B. Pittel. Paths in a Random Digital Tree: Limiting Distributions. *Adv. Appl. Probability*, 18:139–155, 1986.

[11] B. Rais. *Analysis of some Trie Parameters under Probabilistic Models.* PhD thesis, Purdue University, Department of Computer Science, 1992.

[12] B. Rais, P. Jacquet, and W. Szpankowski. A Limiting Distribution for the Depth in PATRICIA Tries. *SIAM Journal on Discrete Mathematics*, 6: 197-213, 1993.

[13] P. Shields. Entropy and Prefixes. *Annals of Probability*, 20:403–409, 1992.

[14] W. Szpankowski. Patricia Tries Again Revisited. *Journal of the ACM*, 37:691–711, 1990.

[15] W. Szpankowski. A Generalized Suffix Tree and its (Un)Expected Asymptotic Behavior. *SIAM Journal on Computing*, 22, 1993.

[16] P. Weiner. Linear Pattern Matching Algorithms. *Proceedings of the 14-th Annual Symposium on Switching and Automata Theory*, 111, 1973.

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399