JIRSS (2004)
Vol. 3, No. 2, pp 139-148

Compact Suffix Trees Resemble PATRICIA Tries: Limiting Distribution of the Depth

Philippe Jacquet[1], **Bonita McVey**[2], **Wojciech Szpankowski**[3*]

[1]INRIA, Rocquencourt, 78153 Le Chesnay Cedex, France.
(Philippe.Jacquet@inria.fr)
[2]Saint Norbert College, De Pere, WI 54115, U.S.A. (mcvebm@mail.snc.edu)
[3]Department of Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A. (spa@cs.purdue.edu)

Abstract. Suffix trees are the most frequently used data structures in algorithms on words. In this paper, we consider the depth of a *compact* suffix tree, also known as the PAT tree, under some simple probabilistic assumptions. For a biased memoryless source, 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 from statistically independent strings. As a result, we show that the limiting distribution for the depth in a PAT tree built over n suffixes is normal.

1 Introduction

Suffix trees have found a wide variety of applications in algorithms on words including: the longest repeated substring [22], squares or

*The work of this author was supported by the NSF Grants CCR-9804760 and CCR-0208709, and NIH grant R01 GM068959-01.

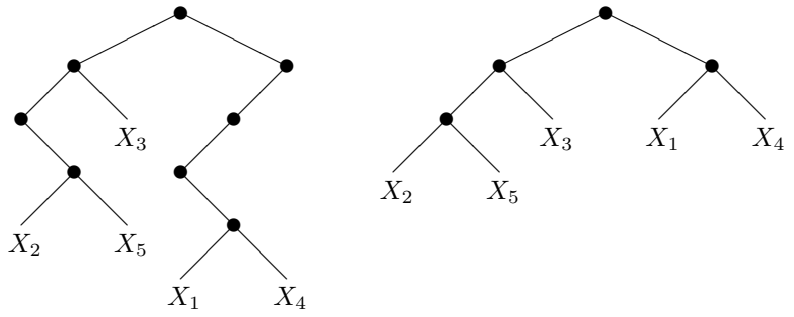
Received: October 2003, Revised: January 2004

Key words and phrases: Digital trees, limiting distribution, Patricia trie, suffix tree.

repetitions in strings [1], string statistics [1], string matching [9, 5], approximate string matching [5], string comparison, compression schemes [11, 12, 23, 24, 25, 26], fast IP addressing schemes [13, 17], biologically significant motif patterns in DNA [5], sequence assembly [9, 5], and so forth (cf. [9, 20, 21]). It is fair to say that suffix trees are the most widely used data structure in algorithms on sequences. Its typical behavior plays a prime role in designing fast and practical algorithms on words. A clear example illustrating the benefits from a probabilistic analysis is given in Chang and Lawler [5], who used some elementary property of a typical behavior of suffix trees to design a superfast algorithm for the approximate string matching problem.

We start with 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 $\Sigma = \{\omega_1, \omega_2, \ldots, \omega_V\}$. We assume that X is generated by a biased memoryless source, that is, $\Pr\{x_j = \omega_i\} = p_i$ 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 also known as the 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 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 (cf. [8, 9, 20]). 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 a single node. The corresponding PAT tree also appears in Figure 1.

Recent 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], but the first complete probabilistic analysis of the height was presented by Devroye, Szpankowski and Rais [6] and Szpankowski [19]. The limiting distribution of the depth in a noncompact suffix tree was analyzed by Jacquet and Szpankowski [10]. In [19] Szpankowski obtained some results involving (a.s.) convergence for the depth, height, and the shortest path length for mixing sources. Also, the external path



suffix tree *PAT tree*

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

length of the noncompact suffix tree was analyzed by Shields [16] and the average size of a suffix tree was established in [10] (cf. [4]). Wyner [24] analyzed the depth for Markov sources. Finally a survey of results for digital trees is given in a book by Gonnet and Baeza-Yates [8] and [20]. It is important to note that previously there were very few known results for the compact suffix tree with the exception of [19] where almost sure convergence of the height was established. In this paper, we analyze the limiting distribution for the depth in a compact suffix tree.

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 *typical depth* is defined to be the depth of a randomly chosen suffix 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.

In the next section we prove our main result that is stated below.

Theorem 2.1. *Consider the PAT tree built over the first n suffixes of a string X generated by a biased memoryless source over a finite alphabet of size V.*

(i) *For large n the average depth* $\mathbf{E}[D_n^{PAT}]$ *attains asymptotically*

$$\mathbf{E}[D_n^{PAT}] = \frac{1}{h}\left(\log n + \gamma + \frac{h_2}{2h}\right) + P_1(\log n) + O\left(n^{-\epsilon}\right), \qquad (2)$$

and the variance $\mathbf{Var}[D_n^{PAT}]$ *of the depth is*

$$\mathbf{Var}[D_n^{PAT}] = \frac{h^2 - h_2}{h^3}\log n + A + P_2(\log n) + O\left(n^{-\epsilon}\right), \qquad (3)$$

where $h = -\sum_{i=1}^{V} p_i \log p_i$ *is the entropy rate,* $h_2 = \sum_{i=1}^{V} p_i \log^2 p_i$, $\gamma = 0.577$ *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 [18].*

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

$$\lim_{n\to\infty} \Pr\{D_n^{PAT} \le \mathbf{E}[D_n^{PAT}] + x\sqrt{\mathbf{Var}[D_n^{PAT}]}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt, \qquad (4)$$

for any fixed x.

Remarks and Observations

(i) *Comparison of the depth in PATRICIA tries and PAT trees.* It appears that the similarities of the trie and the suffix tree carry 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 [15] for the PATRICIA trie, that the PAT tree is, with high probability, well-balanced.

(ii) *Unbiased memoryless source.* Unfortunately, we are unable to extend our results for the depth in a PATRICIA trie to the PAT tree for the unbiased memoryless source in which $p_i = 1/V$ for all $1 \le i \le V$. For the trie, Pittel [14] 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 [10] for suffix trees. Although the proof as described in [14] for the trie is quite simple, the proof for the PATRICIA tree for unbiased memoryless sources is quite complicated, as shown in [15], and we do not know how to extend it to the PAT tree.

(iii) *Markov sources.* We believe one can combine the proof presented in the next section and results of Wyner [24] to extend Theorem 1 to Markov sources.

3 Analysis

In analyzing the depth of the PAT tree, we will make use of the result obtained by Rais, Jacquet and Szpankowski [15] for the depth in a PATRICIA trie, and the result of Jacquet and Szpankowski [10] regarding the limiting distribution for the depth in a (noncompact) 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 strings. This will provide an upper bound for D_n^{PAT}; the limiting distribution of D_n^S proved in [10] is normal with mean $\mathbf{E}[D_n^S]$ and $\mathbf{Var}[D_n^S]$ as expressed by (2) and (3), respectively.

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

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

$$|\Pr\{D_m^{PAT} > k\} - \Pr\{D_m^P > k\}| < \epsilon_m \to 0\,.$$

(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 strings, respectively,

converge to the same distribution. In [15] we have that D_n^P is asymptotically normally distributed with mean $\mathbf{E}[D_n^S]$ and $\mathbf{Var}[D_n^S]$ given in (2) and (3), respectively.

When we have completed these steps, D_n^{PAT} will be bounded by D_n^S and D_n^P which have the same limiting distribution. This will show that the limiting distribution of 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 string in a compact suffix tree is at most equal to the depth of that same string 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 [19] for the computation of the lower bound for the height of a suffix tree. Let $M = \lfloor 2C \log n \rfloor$ where $2C \log n$ is the leading term in the asymptotic height of the suffix tree computed in [6, 19] (in fact, $C = -1/\log(p_1^2 + \ldots + p_V^2)$ for memoryless sources). Then, we choose $Y_i = X_{M(i-1)+1}$ for $i = 1, \ldots, m$ where $m = \lfloor n/M \rfloor = O(\frac{n}{\log n})$. That is, we use every Mth suffix to build a new suffix tree. By choosing the Y_i's in this way, we observe that with high probability the suffixes 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 typical depth of a tree increases when an additional suffix is added to the tree. This is caused by the fact that the typical depth of a tree is defined to be the depth of a randomly chosen suffix as illustrated in (1). However, we can say that the depth of insertion for the Y_i suffix, $D_m^{PAT}(Y_i)$, is stochastically nonincreasing, that is, $D_m^{PAT}(Y_i) \leq_{st} D_n^{PAT}(Y_i)$ $(m \leq n)$ for $i = 1, \ldots, m$ since each Y_i in the tree with m strings is also in the tree with n strings at a depth at least as great as in the tree with m strings. But this also says that for all $k \geq 0$

$$\Pr\{D_m^{PAT}(Y_i) \geq k\} \leq \Pr\{D_n^{PAT}(Y_i) \geq k\}, \tag{5}$$

which leads to the following sequence of steps:

$$\Pr\{D_n^{PAT} \geq k\} \quad \geq \quad \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{m} \Pr\{D_n^{PAT}(X_{M(i-1)+j}) \geq k\}, M \leq n/m$$

$$\begin{aligned} &\overset{(5)}{\geq} \frac{m}{n}\sum_{j=1}^{M}\frac{1}{m}\sum_{i=1}^{m}\Pr\{D_m^{PAT}(Y_i)\geq k\} \\ &= \frac{m}{n}\sum_{j=1}^{M}\Pr\{D_m^{PAT}\geq k\} \\ &\geq \Pr\{D_m^{PAT}\geq k\}+O\left(\frac{\log n}{n}\right). \end{aligned}$$

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

We now present a proof that our PAT tree on the specially chosen m suffixes of X is comparable to a PATRICIA trie on m independent strings. To do this, we construct a second tree whose m strings, Y_i^P for $i = 1, \ldots, m$, are given as follows. The string Y_i^P agrees with the string Y_i on the first M symbols and the remaining symbols are chosen arbitrarily. Obviously, this new tree is a PATRICIA trie since the strings are independent. Thus the limiting distribution D_m^P for the depth of this PATRICIA tree with m independent strings is normal and is given in [15].

Finally, by our choice of M, we know that the $\Pr\{H_n > M\} = O(n^{-\varepsilon}) \to 0$ for $\varepsilon > 0$ as $n \to \infty$ (cf. [6, 19]), where H_n is the height of a suffix tree on n strings. This implies that our compact "suffix" tree on m strings and the PATRICIA tree constructed above are identical with probability tending to 1 (cf. [7]). 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 $k \to \infty$, $(D_k^P - (1/h)\log k)/\sqrt{c_2 \log k}$, where $c_2 = (h^2 - h_2)/h^3$, converges to the standard normal distribution. Since $m = \lfloor n/\lfloor 2C \log n \rfloor \rfloor$ the mean $1/h \log m = 1/h \log n + o(\sqrt{\log n})$, and the variance $c_2 \log m = c_2 \log n - O(\log \log n)$. 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 that for large n,

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

where $\overset{d}{=}$ denotes asymptotic equality in distribution. But D_n^P and D_n^S have the same limiting distribution. Therefore, D_n^{PAT} also has the same limiting distribution which is given explicitly in our theorem. To justify the left-most equality we can use the the multiplicative and

additive theorems[1] in the spirit of Slutsky (cf. [3]). For this we just observe that

$$\begin{aligned}\frac{D_m - c_1 \log n}{\sqrt{c_2 \log n}} &= \frac{D_m - c_1 \log m - c_1 \log\log n}{\sqrt{c_2 \log m}} \left(1 - O\left(\frac{\log\log n}{\log n}\right)\right) \\ &\xrightarrow{d} N(0,1)\end{aligned}$$

where $c_1 = 1/h$. Our proof is now complete.

Acknowledgment

The author thank the referees and Hosam Mahmoud for useful comments that led to a better presentation.

References

[1] Apostolico, A. (1985), The myriad virtue of suffix trees. Springer NATO ASI Ser. F12, March, 85–96.

[2] Apostolico, A. and Szpankowski, W. (1992), self-alignments in words and their applications. Journal of Algorithms, **13**, 446-467.

[3] Billingsley, P. (1986), Probability and Measures. Second Edition, New York: John Wiley & Sons.

[4] Blumer, A., Ehrenfeucht, A., and Haussler, D. (1989), Average sizes of suffix trees and DAWGs. Discrete Applied Mathematics, **24**, 37–45.

[5] Chang, W. and Lawler, E. (1990), Approximate string matching in sublinear expected time. Proceedings of 1990 FOCS, 116–124.

[6] Devroye, L., Szpankowski, W., and Rais, B. (1991), A note on the height of suffix trees. SIAM Journal on Computing, **21**, 48–53.

[7] Devroye, L. and Neininger, R. (2003), Random suffix search trees. Random Structures and Algorithms, **23**, 357-396.

[1] If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{(a.s.)} Y$, then $X_n Y_n \xrightarrow{d} XY$ and $X_n + Y_n \xrightarrow{d} X + Y$.

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

[9] Gusfield, D. (1997), Algorithms on Strings, Trees, and Sequences. Cambridge: Cambridge University Press.

[10] Jacquet, P. and Szpankowski, W. (1994), Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. J. Combinatorial Theory. Ser. A, **66**, 237–269.

[11] Kontoyiannis, I. and Suhov, Y. (1993), Prefix and the entropy rate for long-range sources. In Probability, Statistics, and Optimization (Ed. F.P. Kelly), New York: John Wiley & Sons.

[12] Merhav, N. and Ziv, J. (1997), On the amount of statistical side information required for lossy data compression. IEEE Trans. Information Theory, **43**, 1112–1121.

[13] Nilsson, S. (1996), Radix Sorting & Searching. PhD Thesis, Lund University, 1996.

[14] Pittel, B. (1986), Paths in a random digital tree: Limiting distributions. Adv. Appl. Probability, **18**, 139–155.

[15] Rais, B., Jacquet, P., and Szpankowski, W. (1993), A limiting distribution for the depth in PATRICIA tries. SIAM Journal on Discrete Mathematics, **6**, 197-213.

[16] Shields, P. (1992), Entropy and prefixes. Annals of Probability, **20**, 403–409.

[17] Srinivasan, V. and Varghese, G. (1998), Fast address lookups using controlled prefix expansions. ACM SIGMETRICS'98.

[18] Szpankowski, W. (1990), Patricia tries again revisited. Journal of the ACM, **37**, 691–711.

[19] Szpankowski, W. (1993), A generalized suffix tree and its (un)expected asymptotic behavior. SIAM Journal on Computing, **22**, 1176–1198.

[20] Szpankowski, W. (2001), Average Case Analysis of Algorithms on Sequence. New York: John Wiley & Sons.

[21] Waterman, M. (1995), Introduction to Computational Biology. London: Chapman & Hall.

[22] Weiner, P. (1973), Linear pattern matching algorithms. Proceedings of the 14-th Annual Symposium on Switching and Automata Theory, 1–11.

[23] Wyner, A. and Ziv, J. (1989), Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. IEEE Trans. Information Theory, **35**, 1250–1258.

[24] Wyner, A. J. (1997), The redundancy and distribution of the phrase lengths of the fixed-database Lempel-Ziv algorithm. IEEE Trans. Information Theory, **43**, 1439–1465.

[25] Ziv, J. and Lempel, A. (1977), A universal algorithm for sequential data compression. IEEE Trans. Information Theory, **23**(3), 337–343.

[26] Ziv, J. and Lempel, A. (1978), Compression of individual sequences via variable-rate coding. IEEE Trans. Information Theory, **24**, 530–536.