1992

Compact Suffix Trees Resemble Patricia Tries: Limiting Distribution of Depth

Philippe Jacquet

Bonita Rais

Wojciech Szpankowski

Purdue University, spa@cs.purdue.edu

Report Number:
92-048

https://docs.lib.purdue.edu/cstech/969

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
COMPACT SUFFIX TREES RESEMBLE PATRICIA TRIES: LIMITING DISTRIBUTION OF DEPTH

Philippe Jacquet
Bonita Rais
Wojciech Szpankowski

CSD-TR-92-048
August 1992
COMPACT SUFFIX TREES RESEMBLE PATRICIA TRIES:
LIMITING DISTRIBUTION OF DEPTH

August 5, 1992

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.

*This research was primarily 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.
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]. Heuristic arguments were given by Blumer, Ehrenfeucht and Haussler [3] to show that under certain conditions, the asymptotic expected size of the suffix tree is linear with respect to the number of suffixes stored in the tree. This is proved rigorously in [8]. 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 $X_i = x_ix_{i+1}x_{i+2}...$. 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. An example of a suffix tree for the string $X = 10010011...$ 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,

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

(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 $E D_{n}^{P A T}$ depth of a PAT tree is

$$ED_{n}^{P A T} = \frac{1}{H} \{ \log n + \gamma + \frac{H_{2}}{2H} \} + P_{1}(\log n) + O\left(\frac{1}{n^{1/3}}\right)$$

and the variance $\text{var} D_{n}^{P A T}$ of the depth is

$$\text{var} D_{n}^{P A T} = \frac{H^{2} - H_{2}}{H^{3}} \log n + C + P_{2}(\log n) + O\left(\frac{1}{n^{1/3}}\right)$$

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 $C$ is an explicit constant found in [14].

(ii) The random variable $\frac{D_{n}^{P A T} - ED_{n}^{P A T}}{\sqrt{\text{var} D_{n}^{P A T}}}$ is asymptotically normal with mean zero and variance one, that is,

$$\lim_{n \to \infty} \Pr\{D_{n}^{P A T} \leq ED_{n}^{P A T} + x\sqrt{\text{var} D_{n}^{P A T}}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^{2}/2} dt.$$

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

$$\lim_{n \to \infty} \sup_{x} |\Pr\{D_{n} \leq x\} - e^{-nV - x}| = 0$$

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 $D^\text{PAT}_n \leq D^S_n$ stochastically; that is, for any $x$, we have $\Pr\{D^\text{PAT}_n \geq x\} \leq \Pr\{D^S_n \geq x\}$, where $D^S_n$ is the depth of a noncompact suffix tree with $n$ keys. This will provide an upper bound for $D^\text{PAT}_n$ since the limiting distribution of $D^S_n$ is given in [8] with mean $ED^S_n$ and $\text{var} D^S_n$ 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 $D^\text{PAT}_m$, we show that $D^\text{PAT}_m \leq D^\text{PAT}_n$ stochastically. This provides a lower bound.

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

$$|\Pr\{D^\text{PAT}_m > k\} - \Pr\{D^P_m > k\}| < \epsilon_m.$$  

(iv) Finally, we show for our choice of $m$ that $D^P_m$ and $D^P_n$, the depth of PATRICIA tries with $m$ and $n$ independent keys, respectively, converge to the same distribution. In [12] we have that $D^P_n$ is asymptotically normally distributed with mean $ED^P_n$ and $\text{var} D^P_n$ as given in our theorem.

When we have completed these steps, $D^\text{PAT}_n$ will be bounded by $D^S_n$ and $D^P_n$ which have equivalent limiting distributions. This will show that the limiting distribution of $D^\text{PAT}_n$ is equal to each of them, and will prove that $D^\text{PAT}_n$ is normally distributed.

The first step is easy. Clearly, $D^\text{PAT}_n \leq D^S_n$ 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 $M = [2C\log n]$ where $C\log n$ is the leading term in the asymptotic height of the suffix tree computed in [5, 11, 15] (in fact, $C = 1/\log(p_2^2 + \ldots + p_P^2)$ in the Bernoulli model). Then, we choose $Y_i = X_{M(i-1)+1}$ for
where \( m = \lfloor n/M \rfloor = O\left(\frac{n}{\log n}\right) \). 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 
\[
D^{PAT}_m \leq_{st} D^{PAT}_n
\]
where \( D^{PAT}_m \) 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 
\[
D^{PAT}_m(Y_i) \leq_{st} D^{PAT}_n(Y_i)
\]
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 
\[
\Pr\{D^{PAT}_m(Y_i) \geq k\} \leq \Pr\{D^{PAT}_n(Y_i) \geq k\},
\]
which leads to the following sequence of steps:

\[
\Pr\{D^{PAT}_n \geq k\} = \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{m} \Pr\{D^{PAT}_n(X_{M(i-1)+j}) \geq k\} \\
\geq \frac{m}{n} \sum_{j=1}^{M} \frac{1}{m} \sum_{i=1}^{m} \Pr\{D^{PAT}_m(Y_i) \geq k\} \\
= \frac{m}{n} \sum_{j=1}^{M} \Pr\{D^{PAT}_m \geq k\} \\
\geq \Pr\{D^{PAT}_m \geq k\}.
\]

Thus, \( D^{PAT}_m \) is a lower bound for \( D^{PAT}_n \).

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 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 \( D^P_m \) 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 \( D^P_m \) and \( D^{PAT}_m \) are the same.

Our proof is not yet complete because we cannot equate the limiting distribution of \( D^P_m \) with \( D^S_n \). The problem is that, although \( D^P_m \) and \( D^S_n \) are both normal, \( D^P_m \) has mean and
variance of $O(\log m)$ and $D_n^S$ has mean and variance of $O(\log n)$. However, when $k \to \infty$, $D_n^P$ converges to the normal distribution with mean equivalent to $c_1 \log k$ and variance equivalent to $c_2 \log k$. Since $m = \lfloor n/(2C \log n) \rfloor$ 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 $n$,

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

where $\overset{d}{=}$ 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


