Svffiy Tsfft boe S n utf Spvsdf t

Julien Fayolle
Algorithms Project, INRIA (France)
November 17, 2003
Summary by Pierre Nicodème

Abstract
Using an intricate method, Jacquet and Szpankowski [2] compared the depth of insertion into suffix-trees and tries in the non-uniform Bernoulli model, as well as the average size of suffix-trees and tries under the same model. They proved that the depth of insertion has asymptotically the same probabilistic behaviour in both cases, and that the average sizes of a trie and a suffix-tree built with $n$ keys are asymptotically equivalent. Julien Fayolle uses a simpler combinatorial approach to compare both tree structures. When considering a two-letters alphabet with letters probability $p$ and $q = 1 - p$, he improves the asymptotic estimation for the expectations of external path length and size of the suffix-tree (more specifically, he obtains an asymptotic bound $O(n^{0.85})$ for the difference of the two expectations when $p \in [0.46, 0.54]$). The Lempel–Ziv compression algorithm and its variants use suffix-trees as underlying data-structure.

1. Introduction
We consider a memoryless source over an alphabet $\Sigma = \{0, 1\}$, with $P(0) = p, P(1) = 1 - p = q$ and $p > q$.

Trie. Let $X$ be a finite set of infinite words over $\Sigma$. A trie with input keys $X$ is defined by

$$\text{trie}(X) = \begin{cases} \emptyset & \text{if } |X| = 0, \\ \bullet & \text{if } |X| = 1, \\ \#_{\bullet}, \text{trie}(X\backslash 0), \text{trie}(X\backslash 1) & \text{elsewhere} \end{cases}$$

where the symbol $\bullet$ represents a node and $X\backslash a = \{u : (w \in X \text{ and } w = au)\}$, for $a \in \Sigma$.

Suffix tree. The suffix-tree of $n$ keys over an infinite random string $Y$ is the trie built over the set $X$ of the $n$ first suffixes of $Y$.

Definitions. For any string $\omega \in \Sigma^*$, let $N_\omega$ be the number of keys of $X$ (or first $n$ suffixes of $Y$) whose prefix is $\omega$.

Given a trie $T$, we only consider internal nodes; therefore,

- for any internal node $\nu$ of $T$, if $\omega_\nu$ is the word spelled by reading the labels of the edges of $T$ from the root to $\nu$, we have $N_{\omega_\nu} \geq 2$. We write $\omega_\nu \in T$ in this case;
- reciprocally, if $N_{\omega} = 0$, there is no node in $T$ accessed by reading $\omega$ and if $N_{\omega} = 1$, $\omega$ leads to a leaf; ($\omega \not\in T$ in both cases).
Theorem 1. Asymptotically, for a memoryless source \((p, q)\), the expectation of the external path length \(L\) is:

- if the source is periodic \((\log p/\log q \text{ rational})\)
  \[
  E^{(T_n)}(L) = -\frac{n \log n}{p \log p + q \log q} + Kn + n\epsilon(n) + o(n),
  \]
  where \(\epsilon(n)\) is a periodic function of weak amplitude;
- elsewhere (aperiodic source)
  \[
  E^{(T_n)}(L) = -\frac{n \log n}{p \log p + q \log q} + K + o(n)
  \]

3. Suffix Tree: External Path Length

For any \(\omega\), the probability of occurrence of \(\omega\) as one of the first \(n\) suffixes of the string \(Y\) is independent of the position; therefore \(E^{(S_n)}(N_\omega) = n \times \pi_\omega\).

We similarly have \(E^{(T_n)}(N_\omega) = n \times \pi_\omega = E^{(S_n)}(N_\omega)\).

We want to compute the probability that \(\omega\) occurs once and only once in the string. We use a variation of Guibas and Odlyzko’s method. (See [3] p. 374).

We consider the autocorrelation set \(A_\omega\) of \(\omega\), defined as
\[
A_\omega = \{ h : \omega . h = u.\omega \text{ and } |h| < |\omega| \}.
\]

Let (a) \(F_\omega\), (b) \(T_\omega\) and (c) \(W_\omega\) be respectively the language of texts (a) whose first and lone occurrence of \(\omega\) is at the end of the text (First occurrence in a text), (b) whose concatenation to \(\omega\) do not create a new occurrence of \(\omega\) (Tail following the last occurrence of \(\omega\)) and (c) Without occurrence of \(\omega\). Let \(\sigma\) be any letter of the alphabet \(\Sigma\).

We have two formal equations
\[
W_\omega \sigma = W_\omega + F_\omega - \epsilon \quad \text{and} \quad W_\omega . \omega = F_\omega A_\omega,
\]
where \(\epsilon\) is the empty word (that also belongs to \(A_\omega\)). Products and unions are unambiguous, and we obtain for the weighted generating functions (see Section 1)
\[
W_\omega(z) \times z = W_\omega(z) + F_\omega(z) - 1 \quad \text{and} \quad W_\omega(z) \times \pi_\omega^{|\omega|} = F_\omega(z) \times A_\omega(z).
\]

Solving for \(W_\omega(z)\) and \(F_\omega(z)\), we obtain
\[
F_\omega(z) = \frac{\pi_\omega^{|\omega|}}{\pi_\omega^{|\omega|} + (1 - z) A_\omega(z)}.
\]

Let \(\bar{}\) denote backwards reading of words of a language. For any \(L\), by reading backwards and forwards words, we have a bijection between \(L\) and \(\bar{L}\). This implies (with a memoryless source) that \(L(z) = \bar{L}(z)\). But we have \(W_\omega(z) = \bar{W}_\omega(z) = \bar{W}_{\omega^{|\omega|}}(z) = W_{\omega^{|\omega|}}(z)\) and therefore \(F_\omega(z) = F_{\omega^{|\omega|}}(z)\). This implies that
\[
F_\omega = \bar{F_\omega} \quad \Rightarrow \quad T_{\omega^{|\omega|}}(z) = \frac{\bar{F}_\omega(z)}{\bar{\omega}(z)} = \frac{\bar{F}_{\omega^{|\omega|}}(z)}{\omega(z)} = T_\omega(z) = \frac{F_\omega(z)}{\omega(z)} = \frac{F_{\omega^{|\omega|}}(z)}{\pi_\omega^{|\omega|}}.
\]

We also have \(O_\omega(z) = F_\omega(z) T_\omega(z)\) for the generating function \(O_\omega(z)\) of texts with exactly one match with \(\omega\). Summing up over \(\omega\), we obtain the generating function \(E^{(S)}(z)\) of expectations of
the external path length of a suffix-tree (see Jacquet and Szpankowski [2] for another proof)

\[(2) \quad E_L^{(S)}(z) - \sum_{\omega \in \Sigma^*} E_{N_\omega}^{(S)}(z) = \sum_{\omega \in \Sigma^*} P^{(S)}_{(N_\omega=1)}(z) = \sum_{\omega \in \Sigma^*} O_\omega(z) = \sum_{\omega \in \Sigma^*} \frac{\pi \omega z^{[u]} \cdot (\pi \omega z^{[u]} + (1-z)A_\omega(z))^{-1}}{z}.\]

4. External Path Length, Suffix Tree versus Trie

We consider (the index \(n\) corresponding to \(n\) keys) the differences

\[(3) \quad \Delta_n = E^{(T_n)}(L) - E^{(S_n)}(L) = \sum_{\omega \in \Sigma^*} \delta_{\omega}^{(n)} = \sum_{\omega \in \Sigma^*} P^{(T_n)}(N_\omega = 1) - P^{(S_n)}(N_\omega = 1),\]

(since \(E^{(T_n)}(N_\omega) = E^{(S_n)}(N_\omega)\) for all \(\omega\)). We will prove that \(\Delta_n = O(n^{0.85})\) for \(0.5 < p < 0.54\).

The minimum \(\mu_n\) of the fillup levels of both trees is \(\alpha \log(n)\) for a given \(\alpha > 0\) with probability one; all the following will be conditioned by the fact that \(\mu_n = \alpha \log(n)\). With \(|\omega| < \mu_n\), we have \(\delta_{\omega}^{(n)} = 0\) and when \(|\omega| \geq \mu_n\), asymptotically, \(\pi_\omega = o(1)\).

4.1. Asymptotic contribution of the trie. As we have

\[(4) \quad P^{(T_n)}(N_\omega = 1) = n\pi_\omega \times (1 - \pi_\omega)^{n-1} \times n\pi_\omega \times e^{-n\pi_\omega}.\]

4.2. Asymptotic contribution of the suffix-tree. Each function \(O_\omega(z)\) in Equation 2 is a rational function with dominant pole \(p_\omega\). We begin by isolating the dominant poles of these functions that give the asymptotic behaviour of the terms corresponding to the suffix-tree in \(\Delta_n\).

Lemma 1. For \(1 < R < 1/p\) and \(|\omega| \geq \mu_n\), the set of poles of the functions \(O_\omega(z)\) contained in the disk centered at the origin and of radius \(R\) is exactly \(\{p_\omega; \omega \in \Sigma^*, |\omega| \geq \mu_n\}\).

Proof. The proof is based on an application of the Rouché theorem; let \(f(z) = \pi_\omega z^{[u]}\) and \(g(z) = (1-z)A_\omega(z)\). We are above the level \(\mu_n\) and therefore, for \(|z| < 1/p\), we have \(|f(z)| = o(1)\). Let \(d\) be the smallest period of \(\omega\).

- If \(d \leq |\omega|/2\), we have \(\omega = u^r v, (|v| < |u| = d \text{ and } v \text{ a prefix of } u)\), and the second smallest period is at least \(|\omega|/2\) (a consequence of the Wilf theorem). This gives (omitting the subscript \(\omega\))

\[
A(z) = 1 + S(z) + T(z), \quad \text{with} \quad S(z) = \pi_{u^r v^{[u]}} + \cdots + (\pi_{u^r v^{[u]}})^r,
\]

where \(T(z)\) is a polynomial of lowest degree \(\geq \mu_n/2\); this implies \(|T(z)| = o(1)|z| < 1/p\) and

\[
|A(z)| = \left| \frac{1}{1 - \pi_{u^r v^{[u]}}} \right| + o(1).
\]

For all \(d \geq 2\), this implies (up to negligible terms)

\[
|A(z)| \geq \frac{1}{1 + \pi_{u^r v^{[u]}}} \geq \frac{1}{1 + p|z|} \quad \text{for} \quad |z| < \frac{1}{p}.
\]

We have the same lower bound for \(d = 1\).

- If \(d > |\omega|/2\), we have \(S(z) = 0\) and \(|A(z)| = 1 + o(1)\) for \(|z| < 1/p\).

Therefore, for any \(R < 1/p\) (and we choose in the following \(R > 1\)), there exists a number \(N\) such that for \(n > N\), on the circle \(|z| = R\), we have \(|A_\omega(z)| \geq k\) for a given \(k > 0\) and for all \(\omega\) such that \(|\omega| > \mu_n\); this implies \(|f| \geq |g|\) over this circle. Moreover \(f\) and \(g\) are analytic everywhere, which implies that \(f + g\) has as many zeros as \(g\) inside the disk \(D_R = \{z; |z| < R\}\), for any \(\omega\). The polynomial \(A(z)\) has no roots inside the disk \(|z| < R\) and therefore \(f(z) + g(z)\) has only one root inside the disk \(D_R\). \(\square\)
For each $\omega$, we compute
\[ \alpha_n^{(\omega)} = \left[ z^n \right] \Omega_n(z) = \text{Res} \left( \frac{\Omega_n(z)}{z^{n+1}}, 0 \right) = I(C_R) - \text{Res} \left( \frac{\Omega_n(z)}{z^{n+1}}, \rho_\omega \right), \]
where $I(C_R) = \int_{C_R} \frac{\Omega_n(z)}{z^{n+1}} dz$, and $C_R$ is the circle $|z| = R$. Considering
\[ \Omega_n(z) = \left( \frac{\pi \omega z^{\omega|\omega|}}{(\pi \omega z^{\omega|\omega|} + (1 - z)A_n(z))^2} \right) \]
for $|\omega| > \mu_n$ and $|z| = R$ ($R \in ]1, 1/p[)$ we have $\pi \omega z^{\omega|\omega|} = o(1)$ and $|(1 - z)A_n(z)| \geq (R - 1)\kappa$. Therefore $I(C_R) = O(R^{-n})$.

By bootstrapping, we have $\rho_\omega = 1 + \pi_\omega/A_n(1) + o(\pi_\omega)$. An expansion of the denominator of $\Omega_n$ in a neighborhood of $\rho_\omega$ gives
\[ \mathbf{P}^{(S_n)}(N_\omega = 1) = \alpha_n^{(\omega)} = n\pi_\omega e^{-\frac{n\pi_\omega}{\lambda_{\omega,\omega}}} + O \left( n\pi_\omega^2e^{-\frac{n\pi_\omega}{\lambda_{\omega,\omega}}} \right) + O \left( |\omega|n\pi_\omega^2 \right) + C \left( \frac{1}{R} \right)^n. \]

Plugging Equations 4 and 5 into Equation 3 gives
\[ \Delta_n = \sum_{\omega \in \Sigma^*, |\omega| > \mu_n} \delta_n^{(\omega)} \text{ with } \delta_n^{(\omega)} = \sum_{\omega \in \Sigma^*, |\omega| > \mu_n} n\pi_\omega \left( e^{-nm_\omega/A_n(1)} - e^{-nm_\omega} \right). \]

4.3. Bounding $\Delta_n$ by partitioning the motifs $\omega$. The aim of this section is to prove the validity of each entry of the following table (where $C_p = \log 2/\log(1/p)$ and $\Omega_5^{(n)}, \Omega_{4,p}^{(n)}, \Omega_{3,p}^{(n)},$ and $\Omega_4^{(n)}$ respectively are the set of short, intermediate periodic, intermediate aperiodic and long motifs when the number of input keys is $n$).

| Short patterns $|\omega| < \frac{3}{6} \log_1/n$ | Intermediate patterns $\frac{15\log_1/n}{n} < |\omega|$ | Long patterns $1.5 \log_1/n < |\omega|$ |
|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| $\Delta_n^{(S)} = \sum_{\omega \in \Omega_5^{(n)}} \delta_n^{(\omega)} = o(1)$ | $\Delta_n^{(I,p)} = \sum_{\omega \in \Omega_{4,p}^{(n)}} \delta_n^{(\omega)} = O(n^{0.75}C_p \log n)$ | $\Delta_n^{(I)} = \sum_{\omega \in \Omega_4^{(n)}} \delta_n^{(\omega)} = O \left( \sqrt{n} \right)$ |
| $\mu_n < |\omega|$| periodic ($A_n(1) \geq 1 + 2^{-|\omega|/2}$) | aperiodic ($A_n(1) < 1 + 2^{-|\omega|/2}$) |

4.3.1. Short patterns. For these patterns, we have
\[ n\pi_\omega > nq_{1/\omega}^{5 \log_1/n} = n^{1 - 5} = n^{1/6} \to \infty \quad \text{and} \quad \left| \Omega_s^{(n)} \right| = O(n^\alpha \log n) \text{ where } \alpha = \frac{5 \log 2}{6 \log(1/q)}. \]

Therefore $\Delta_n^{(S)}$ behaves as $n^\alpha \log n \times e^{-n^{1/6}}$ as $n$ tends to infinity and is $o(1)$.

4.3.2. Long patterns. In this case, let $k_1(n) = 1.5 \log_1/n$; we have
\[ n\pi_\omega \leq np_{k_1(n)} = np_{1.5 \log_1/n} = n^{-0.5} \to 0. \]

Expanding $\delta_n^{(\omega)}$ for small $n\pi_\omega$ gives $\delta_n^{(\omega)} \sim (n\pi_\omega)^2 (1 - 1/A_n(1))$. Therefore we have
\[ \Delta_n^{(I)} = \sum_{k \geq k_1(n)} \sum_{\omega \in \Sigma^k} n^2 \pi_\omega \left( 1 - \frac{1}{A_n(1)} \right) \quad \text{and} \quad \sum_{\omega \in \Sigma^k} \pi_\omega = \sum_{0 \leq i \leq k} \left( \frac{k}{i} \right) p^{2i}q^{(k-i)} = (p^2 + q^2)^k < p^k. \]
This implies $\Delta_n^{(1)} = O(n^2 p k_i(n)) = O(\sqrt{n})$.

4.3.3. Intermediate patterns. Julien Fayolle proves in [1] that $\sum_{\omega \in \Sigma^k} A_\omega(1) = 2^k + k - 1$. He defines the set of intermediate periodic patterns as

$$\Omega_{i,p}^{(n)} = \{ \omega : k_5(n) < |\omega| \leq k_1(n), A_\omega(1) \geq 1 + 2^{-k/2} \},$$

where $k_5(n) = 5/6 \log_2 n$ (in the uniform case these patterns verify $d < |\omega|/2$). He also proves that $|\Omega_{i,p}^{(n)}| \leq k 2^{k/2}$.

Summing up for the intermediate periodic patterns, we obtain

$$\Delta_n^{i,p} = \sum_{\omega \in \Omega_{i,p}^{(n)}} \delta_\omega^{(n)} < K \sum_{k=k_5(n)}^{k_1(n)} k 2^{k/2} = O\left(\log n \times e^{1.5 \log \log \log n} \right) = O(n^{0.85}) \quad \text{for} \quad p < 0.54.$$}

The set $\Omega_{i,a}^{(n)}$ of intermediate aperiodic patterns is the complementary set of $\Omega_{i,p}^{(n)}$, inside the bounds $k_5(n) < |\omega| < k_1(n)$. We therefore have $1/A_\omega(1) \geq 1/(1 + 2^{-|\omega|/2}) \geq 1 - 2^{-|\omega|/2}$ and for $|\omega| = k$

$$\delta_\omega^{(n)} \leq n \pi \omega \left( e^{n \pi 2^{-k/2} - 1} \right).$$

We also have for $\omega \in \Omega_{i,a}^{(n)}$

$$n \pi \omega 2^{-|\omega|/2} < n p k_5(n) 2^{-k_5(n)/2} \rightarrow 0 \quad \text{for} \quad \frac{5 \log(p/2)}{6 \log(1/q)} + 1 < 0 \quad \text{or} \quad p < p_0 = 0.5469.$$}

By expanding the exponential in the neighborhood of zero, we get $\delta_\omega^{(n)} < (n \pi \omega)^2 e^{-n \pi 2^{-|\omega|/2}}$, which gives (remarking that $x^2 e^{-x}$ is bounded on $\mathbb{R}^+$)

$$\Delta_n^{i,a} \leq \sum_{k=k_5(n)}^{k_1(n)} K' k^2 2^{-k/2} = O\left(\left( e^{1.5 \log \log \log n} \right) \right) = O(n^{0.85}) \quad \text{for} \quad p < p_0.$$}

Remark that we also have $\Delta_n^{i,p} = O(n^{0.85})$ for $p < p_0$.

4.4. End result. Summarizing the preceding results gives

**Theorem 2.** For a suffix-tree with $n$ keys, we have asymptotically for $p \leq 0.54$

$$E(L_n^{(S)}) = \frac{n \log n}{p \log p + q \log q} + (K + \epsilon(n)) n + O\left( n^{0.85} \right),$$

where $L_n^{(S)}$ is the external path length of the tree and $\epsilon(n)$ is a periodic function of small amplitude.

The same method applies for analysis of the asymptotic expectation of size.

**Bibliography**