Asymptotic Properties of Suffix Trees

Analysis of height and feasible path length

Overview by

Ivan Kazmenko

Saint Petersburg State University, Russia

Original paper by

Wojciech Szpankowski

Purdue University, USA

Plan of the Talk

1. Suffix Trees: Construction

2. Depth of Insertion in a Suffix Tree

3. Height and Shortest Feasible Path in a Suffix Tree

4. Proof Techniques

Definitions

- Σ is a finite alphabet, $|\Sigma| = V$
- $\{X_k\}_{k=1}^{\infty}$ is a stationary ergodic sequence of symbols generated from Σ
- $X_m^n = (X_m, ..., X_n)$ for $m < n$ is a partial sequence

The Problem: Construction

Consider a digital tree built in the following way:

Step 0. At the beginning, the tree consists of its root only.

Step 1. Consider a tree $\mathcal{T}_n$ built for the partial sequence $X_1^n = (X_1, ..., X_n)$.

Step 2. Set current vertex to root.

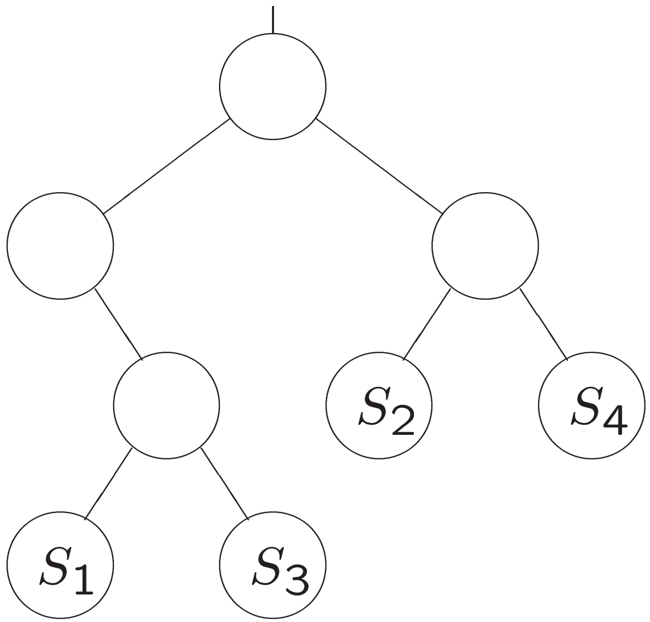
Step 3. Starting with $j = n + 1$, we either

(A) move by the edge marked by X_j from the current vertex if it exists thus changing the current vertex and increase j by 1, or

(B) construct a new edge marked with symbol X_j from the current vertex to a new vertex marked with our suffix X_{n+1}^{∞} and proceed to Step 1 with n increased by 1 otherwise

Example

Tree with 4 inserted suffixes.



Let $X_1^{10} = (0, 1, 0, 1, 1, 0, 1, 1, 1, 0)$.

$S_1 = 0101101110$

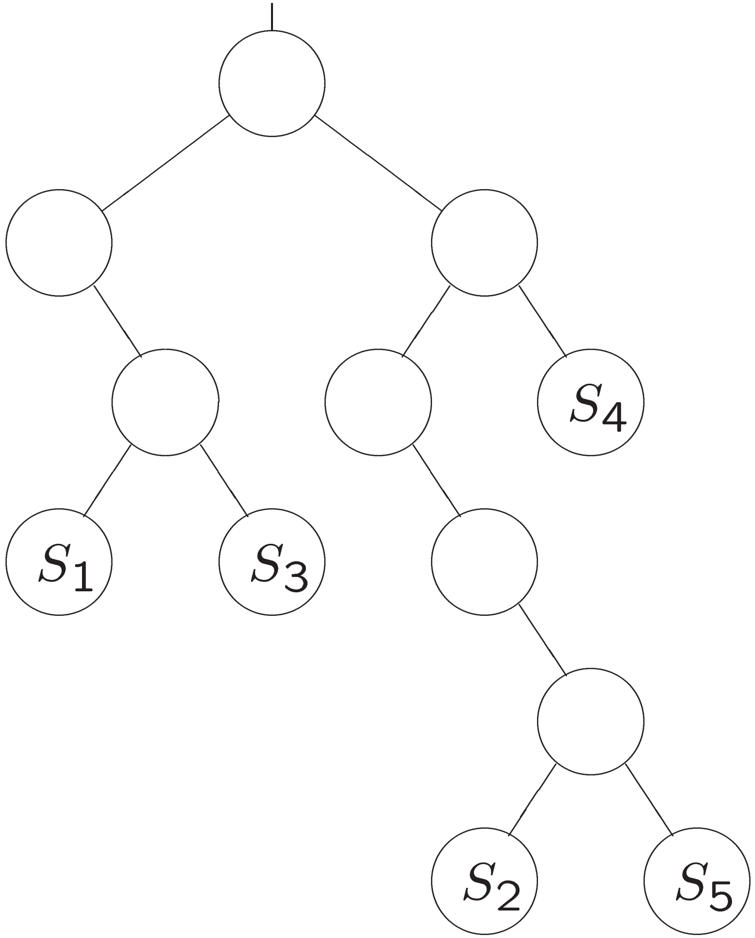
$S_2 = 101101110$

$S_3 = 01101110$

$S_4 = 1101110$

Example

Fifth suffix insertion.



Let $X_1^{10} = (0, 1, 0, 1, 1, 0, 1, 1, 1, 0)$.

$S_1 = 0101101110$

$S_2 = 101101110$

$S_3 = 01101110$

$S_4 = 1101110$

$S_5 = 101110$

The Problem: Questions

- What is the typical height of $\mathcal{T}_n$?

- What is the typical difference $j - n$ when Step 3 is finished?

- What is the typical minimal possible difference $j - n$ at the end of Step 3 for the tree $\mathcal{T}_n$?

  Note that $j - n$ is the number of case (A) occurences during a single Step 3.

More Definitions

- Σ is a finite alphabet, $|\Sigma| = V$
- $\{X_k\}_{k=1}^{\infty}$ is a stationary ergodic sequence of symbols generated from Σ
- $X_m^n = (X_m, ..., X_n)$ for $m < n$ is a partial sequence
- $P(X_1^n) = Pr\{X_k = x_k, 1 \leqslant k \leqslant n, x_k \in \Sigma\}$ is nth order probability distribution
- $h = \lim_{n \to \infty} \frac{E\{-\log P(X_1^n)\}}{n}$ is the entropy of $\{X_k\}$

  It is known that $h \leqslant \log V$.

Parameter L_n

- L_n is the smallest integer $L > 0$ such that $X_m^{m+L-1} \neq X_{n+1}^{n+L}$ for all $1 \leqslant m \leqslant n$.

Example:

Let $X_1^{10} = (0, 1, 0, 1, 1, 0, 1, 1, 1, 0)$.
Here $L_1 = 1$, $L_2 = 3$, $L_3 = 2$, and $L_4 = 5$ since $X_5^8 = X_2^5 = (1, 0, 1, 1)$ and therefore $L_4 > 4$:

$$(0, \underbrace{1, 0, 1, \color{red}{1}}\color{red}{, 0, 1, 1}, 1, 0).$$

Mixing Condition

Let F_m^n be a σ-field generated by $\{X_k\}_{k=m}^n$ for $m \leqslant n$.

$\{X_k\}$ satisfies the mixing condition $\iff$ there exist constants $0 < c_1 \leqslant c_2$ and an integer d such that for all $A \in F_{-\infty}^m$, $B \in F_{m+d}^{\infty}$ and $-\infty \leqslant m \leqslant m + d \leqslant n$ the following condition is true:

$$c_1 Pr\{A\} Pr\{B\} \leqslant Pr\{AB\} \leqslant c_2 Pr\{A\} Pr\{B\}.$$

Strong α-Mixing Condition

Let α be a function of d such that $\alpha(d) \xrightarrow[d\to\infty]{} 0$.

$\{X_k\}$ satisfies the strong α-mixing condition $\iff$ for all $A \in F_{-\infty}^{m}$, $B \in F_{m+d}^{\infty}$ and $-\infty \leqslant m \leqslant m + d \leqslant n$ the following condition is true:

$$(1 - \alpha(d))Pr\{A\}Pr\{B\} \leqslant Pr\{AB\} \leqslant (1 + \alpha(d))Pr\{A\}Pr\{B\}.$$

Parameters h_1 and h_2

$$h_1 = \lim_{n\to\infty} \frac{\max\{\log P^{-1}(X_1^n),\ P(X_1^n)>0\}}{n} = \lim_{n\to\infty} \frac{\log(1/\min\{P(X_1^n),\ P(X_1^n)>0\})}{n}$$

$$h_2 = \lim_{n\to\infty} \frac{\log(E\{P(X_1^n)\})^{-1}}{2n} = \lim_{n\to\infty} \frac{\log(\sum_{X_1^n} P^2(X_1^n))^{-1}}{2n}$$

The relationship with entropy h is as follows:

$$0 \leqslant h_2 \leqslant h \leqslant h_1.$$

Example: Bernoulli Model

Assume that symbols X_i are generated indepenently,
and ith symbol is generated according to the probability p_i.
Thus, $h = \sum_{i=1}^{V} p_i \log(p_i^{-1})$, $h_1 = \log(1/p_{min})$ and $h_2 = 2\log(1/P)$
where $p_{min} = \min_{1 \leqslant i \leqslant V} \{p_i\}$ is the probability
of least probable symbol occurence
and $P = \sum_{i=1}^{V} p_i^2$ can be interpreted as a probability of a match
between any two symbols.

Theorem 1

Consider stationary ergodic sequence $\{X_k\}_{k=-\infty}^{\infty}$.

- Assume strong α-mixing condition
- Let $h_1 < \infty$ and $h_2 > 0$
- $(*)$ $\exists \rho : 0 < \rho < 1, \exists \beta$ such that $\alpha(d) = O(d^{\beta}\rho^{d})$ for $d \to \infty$

  Then

  (1) $\liminf\limits_{n\to\infty} \frac{L_n}{\log n} = \frac{1}{h_1}$ (a.s.) ,

  (2) $\limsup\limits_{n\to\infty} \frac{L_n}{\log n} = \frac{1}{h_2}$ (a.s.) .

Is the Condition $(*)$ Restrictive?

- In Bernoulli model, $\alpha(d) = 0$ because of independence of X_k.
- If the sequence $\{X_k\}$ is Markovian, $\alpha(d)$ decays exponentially fast
- In general, statement (1) of Theorem 1 does not hold without the $(*)$ condition

Depth in a Suffix Tree

Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of symbols from Σ.
Let $\mathcal{T}_n$ be a suffix tree constructed from the first n suffixes of $\{X_k\}$.

- mth depth $L_n(m)$ is the depth of the ith suffix in $\mathcal{T}_n$;
  note that $L_n = L_{n+1}(n+1)$
- Average depth D_n is the depth of a randomly selected suffix,
  that is, $D_n = \frac{1}{n} \sum_{m=1}^{n} L_n(m)$

Height and Shortest Feasible Path

- Height H_n is the length of the longest path in $\mathcal{T}_n$; $H_n = \max\limits_{1 \leqslant m \leqslant n} \{L_n(m)\}$.

- Available node is a node which does not belong to $\mathcal{T}_n$ but its predecessor does, that is, a node that could be inserted in $\mathcal{T}_{n+1}$ at the next insertion.

- Shortest feasible path s_n is the length of the shortest path from the root to an available node.

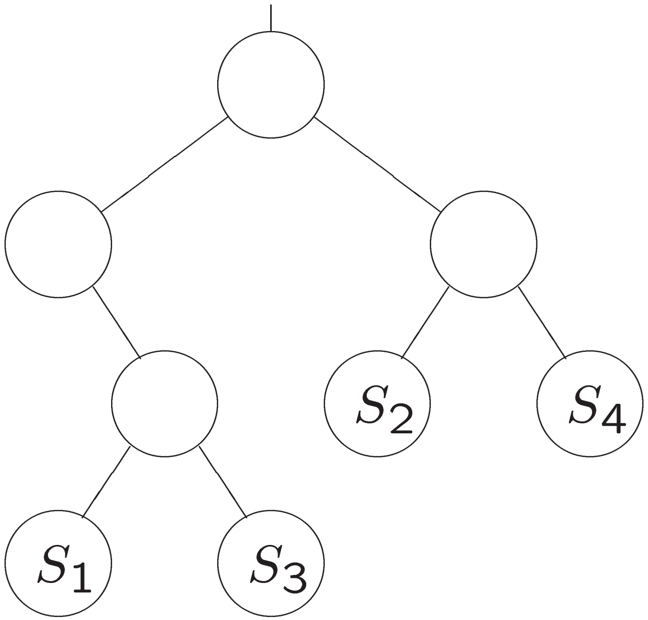
Self-alignment

Let the suffix tree $\mathcal{T}_n$ be built from the suffixes $S_1, ..., S_n$.
Self-alignment $C_{i,j}$ is the length of the longest common prefix of S_i and S_j.

Relation to other suffix tree parameters:

- $L_n(m) = \max_{1 \leqslant k \leqslant n, k \neq m} \{C_{k,m}\} + 1$
- $H_n = \max_{1 \leqslant i < j \leqslant n} \{C_{i,j}\} + 1$
- $L_n = \max_{1 \leqslant m \leqslant n} \{C_{m,n+1}\} + 1$

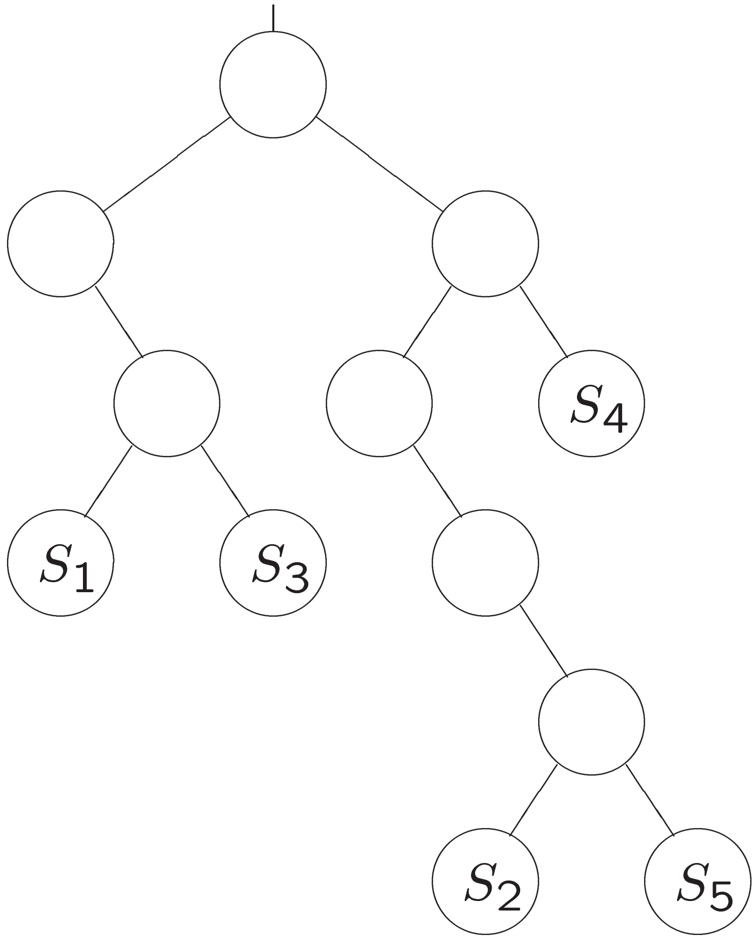
Example



$S_1 = 0101101110$
$S_2 = 101101110$
$S_3 = 01101110$
$S_4 = 1101110$

Let $X_1^{10} = (0, 1, 0, 1, 1, 0, 1, 1, 1, 0)$.
Consider suffix tree $\mathcal{T}_4$ built from first 4 suffixes.
$L_4(1) = 3$, $L_4(2) = 2$, $L_4(3) = 3$, $L_4(4) = 2$.
$H_4 = 3$, $s_4 = 2$.
But $L_4 = L_5(5) = 5$.

Example



$S_1 = 0101101110$
$S_2 = 101101110$
$S_3 = 01101110$
$S_4 = 1101110$
$S_5 = 101110$

But $L_4 = L_5(5) = 5$.
$H_5 = 5$, and $s_5 = 2 = s_4$.

Theorem 2

Consider stationary ergodic sequence $\{X_k\}_{k=1}^{\infty}$.

- Assume strong α-mixing condition
- Let $h_1 < \infty$ and $h_2 > 0$

  Then

  (1) $\lim_{n\to\infty} \frac{s_n}{\log n} = \frac{1}{h_1}$ (a.s.) when $(*)$ holds,

  (2) $\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{1}{h_2}$ (a.s.) when $\alpha(d)$ satisfies the following:

  $\sum_{d=0}^{\infty} \alpha^2(d) < \infty$.

Proof of Theorem 1 by Theorem 2

(1):

$$\limsup_{n\to\infty} \frac{L_n}{\log n} \leqslant \lim_{n\to\infty} \frac{H_n}{\log n} \text{ (a.s.):}$$

by definition: $L_n \leqslant H_n$.

Proof of Theorem 1 by Theorem 2

(1):

$\limsup_{n\to\infty} \frac{L_n}{\log n} \leqslant \lim_{n\to\infty} \frac{H_n}{\log n}$ (a.s.):

by definition: $L_n \leqslant H_n$.

$\limsup_{n\to\infty} \frac{L_n}{\log n} \geqslant \lim_{n\to\infty} \frac{H_n}{\log n}$ (a.s.):

Note that H_n is a non-decreasing sequence;

$L_n = H_n$ a.s. when $H_{n+1} > H_n$, and that occurs infinitely often since $H_n \to \infty$ and $\{X_k\}$ is an ergodic sequence, so

$$Pr\{L_n = H_n \ i.o.\} = 1$$

and there exists a subsequence $n_k \to \infty$ such that $L_{n_k} = H_{n_k}$.

Proof of Theorem 1 by Theorem 2

(1):

$$\limsup_{n\to\infty} \frac{L_n}{\log n} \leqslant \lim_{n\to\infty} \frac{H_n}{\log n} \text{ (a.s.):}$$

by definition: $L_n \leqslant H_n$.

$$\limsup_{n\to\infty} \frac{L_n}{\log n} \geqslant \lim_{n\to\infty} \frac{H_n}{\log n} \text{ (a.s.):}$$

Note that H_n is a non-decreasing sequence;

$L_n = H_n$ a.s. when $H_{n+1} > H_n$, and that occurs infinitely often since $H_n \to \infty$ and $\{X_k\}$ is an ergodic sequence, so

$$Pr\{L_n = H_n \ i.o.\} = 1$$

and there exists a subsequence $n_k \to \infty$ such that $L_{n_k} = H_{n_k}$.

(2) can be proved in a similar way:

s_n is a non-decreasing sequence also.

Techniques: String-Ruler Approach

- Summary: The correlation between different substrings is measured using another string ω called a string-ruler.

- Example:
  How to find the longest common prefix of two independent strings $\{X_k(1)\}_{k=1}^{\infty}$ and $\{X_k(2)\}_{k=1}^{\infty}$?
  Let its length be $C_{1,2}$.
  $C_{1,2} \geqslant k \iff \exists \omega$ of length k: $X_1^k(1) = \omega = X_1^k(2)$.
  We then construct a set $\mathcal{W}_k = \{\omega \in \Sigma^k : |\omega| = k\}$ and estimate the probabilities $P(\omega_k) = P(X_{m+1}^{m+k} = \omega_k)$ for a fixed position m in our sequence $\{X_k\}$.

Techniques: Second Moment Method

- Summary: Second Moment Method by Chung and Erdös:
  For a sequence of events A_i we have
  $$Pr\{\bigcup_{i=1}^{n} A_i\} \geqslant \frac{(\sum_{i=1}^{n} Pr\{A_i\})^2}{\sum_{i=1}^{n} Pr\{A_i\} + \sum_{i \neq j} Pr\{A_i \cap A_j\}}.$$

- Application:

  We then set $A_{i,j} = \{C_{i,j} \geqslant k\}$.

Techniques: Second Moment Method

- Reasoning:

  Markov's Inequality:

  $Pr\{X \geqslant t\} \leqslant \frac{E\{X\}}{t}$.

  Chebyshev's Inequality:

  $Pr\{|X - E\{X\}| \geqslant t\} \leqslant \frac{Var\{X\}}{t^2}$.

- Trivial Results:

  First Moment Method:

  For integer-valued nonnegative random variable X

  $Pr\{X > 0\} \leqslant E\{X\}$.

  Second Moment Method (Chebyshev):

  $Pr\{X = 0\} \leqslant \frac{Var\{X\}}{(E\{X\})^2}$.

References

1. Wojciech Szpankowski, Asymptotic properties of data compression and suffix trees, IEEE Transactions on Information Theory 39 (1993), no. 5, pp. 1647-1659.

2. Wojciech Szpankowski, Average case analysis of algorithms on sequences; available online as http://www.cs.purdue.edu/homes/spa/book.html.