ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

Work based on an idea of Philippe,

begun with him at the end of 2010

ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

Work based on an idea of Philippe,

begun with him at the end of 2010

Dedicated to his memory.

ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

Work based on an idea of Philippe,

begun with him at the end of 2010

Dedicated to his memory.

Journées ALEA, Mars 2013

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.

- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.

- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.
- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

This realistic analysis is successful for two other types of trees :

- Tries and BST, when they are built on general sources
- Why not DST, since it is a mixing of these two structures?

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.
- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

This realistic analysis is successful for two other types of trees :

- Tries and BST, when they are built on general sources
- Why not DST, since it is a mixing of these two structures?

This talk : Analysis of Digital Search Trees

when they are built on words emitted by a general source.

I. Trees.

Description of the DST structure (I)

On the alphabet $\Sigma := \{a, b\}$, consider the sequence of six words

$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

Description of the DST structure (I)

On the alphabet $\Sigma := \{a, b\}$, consider the sequence of six words

$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

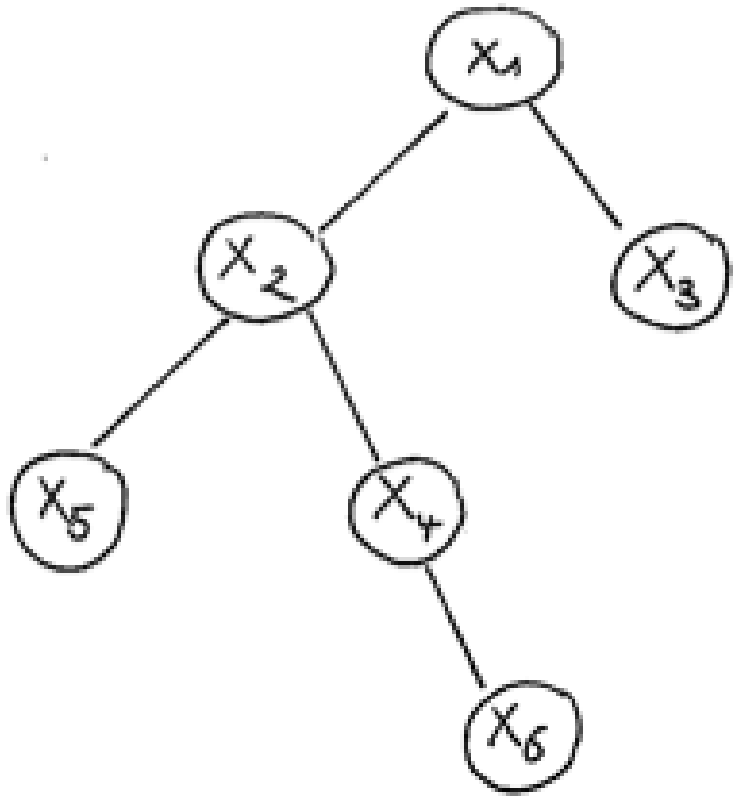
and the digital search tree built on this sequence

Description of the DST structure (I)

On the alphabet $\Sigma := \{a, b\}$, consider the sequence of six words

$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

and the digital search tree built on this sequence



Description of the DST structure (II)

$\mathcal{X}$ = an ordered sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

$$\mathcal{X} := \{X_1, X_2, \ldots, X_n\}$$

Description of the DST structure (II)

$\mathcal{X}$ = an ordered sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

$$\mathcal{X} := \{X_1, X_2, \ldots, X_n\}$$

The first word of the sequence $\mathcal{X}$ is placed at the root.

Root [DST ($\mathcal{X}$)] := First ($\mathcal{X}$).

Description of the DST structure (II)

$\mathcal{X}$ = an ordered sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

$$\mathcal{X} := \{X_1, X_2, \ldots, X_n\}$$

The first word of the sequence $\mathcal{X}$ is placed at the root.

$$\text{Root [DST } (\mathcal{X})] := \text{First } (\mathcal{X}).$$

There are two subtrees built with the sequence $\mathcal{Y} := \mathcal{X} \setminus \{\text{First}(\mathcal{X})\}$,

Description of the DST structure (II)

$\mathcal{X}$ = an ordered sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

$$\mathcal{X} := \{X_1, X_2, \ldots, X_n\}$$

The first word of the sequence $\mathcal{X}$ is placed at the root.

$$\text{Root [DST } (\mathcal{X})] := \text{First } (\mathcal{X}).$$

There are two subtrees built with the sequence $\mathcal{Y} := \mathcal{X} \setminus \{\text{First}(\mathcal{X})\}$,

- The left subtree contains the words of $\mathcal{Y}$ which begin with a

  $\mathcal{Y}_{(a)}$ = the subsequence of $\mathcal{Y}$ formed with words which begin with a

$$\text{Left [DST } (\mathcal{X})] := \text{DST } (\mathcal{Y}_{(a)}).$$

Description of the DST structure (II)

$\mathcal{X}$ = an ordered sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

$$\mathcal{X} := \{X_1, X_2, \ldots, X_n\}$$

The first word of the sequence $\mathcal{X}$ is placed at the root.

$$\text{Root [DST } (\mathcal{X})] := \text{First } (\mathcal{X}).$$

There are two subtrees built with the sequence $\mathcal{Y} := \mathcal{X} \setminus \{\text{First}(\mathcal{X})\}$,

- The left subtree contains the words of $\mathcal{Y}$ which begin with a

  $\mathcal{Y}_{(a)}$ = the subsequence of $\mathcal{Y}$ formed with words which begin with a

  $$\text{Left [DST } (\mathcal{X})] := \text{DST } (\mathcal{Y}_{(a)}).$$

- The right subtree contains the words of $\mathcal{Y}$ which begin with b.

  $\mathcal{Y}_{(b)}$ = the subsequence of $\mathcal{Y}$ formed with words which begin with b

  $$\text{Right [DST } (\mathcal{X})] := \text{DST } (\mathcal{Y}_{(b)}).$$

The three tree structures and their (internal or external) path length

For $\Sigma := \{a, b\}$, consider the case of the sequence of six words

$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

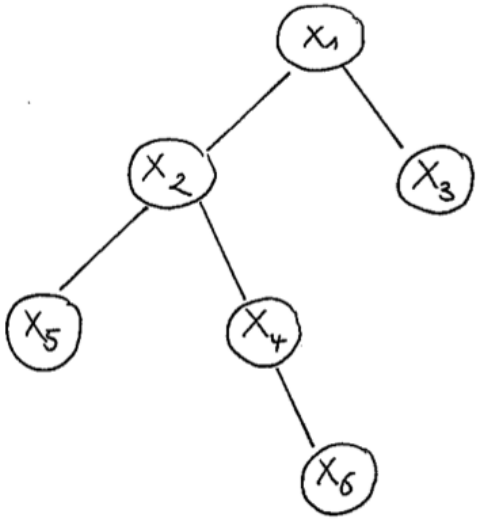
DST Trie Symbol-BST

The three tree structures and their (internal or external) path length

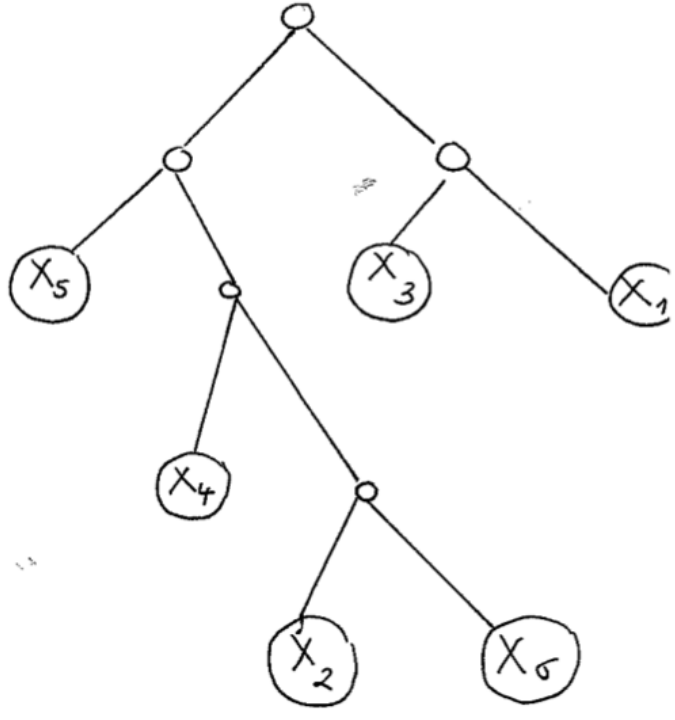
For $\Sigma := \{a, b\}$, consider the case of the sequence of six words

$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

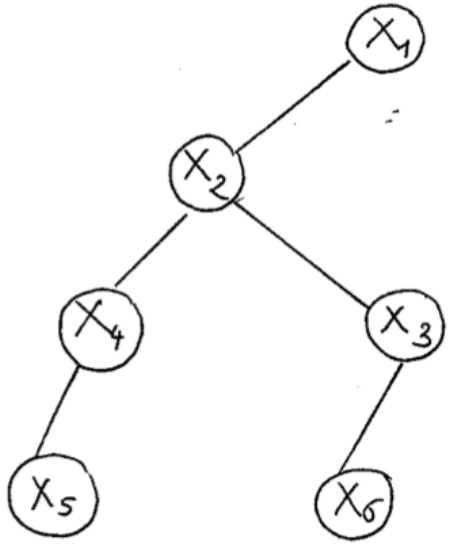
DST



Trie



Symbol-BST

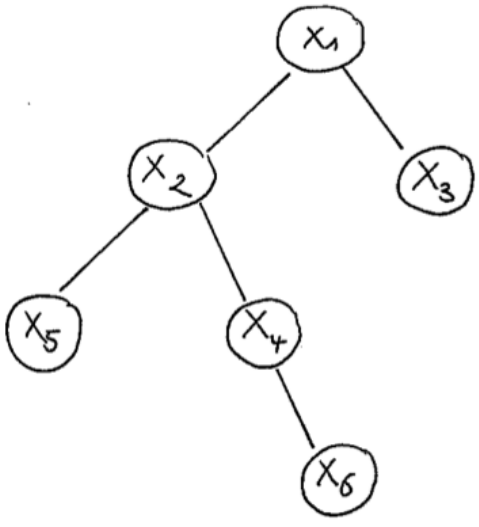


The three tree structures and their (internal or external) path length

For $\Sigma := \{a, b\}$, consider the case of the sequence of six words

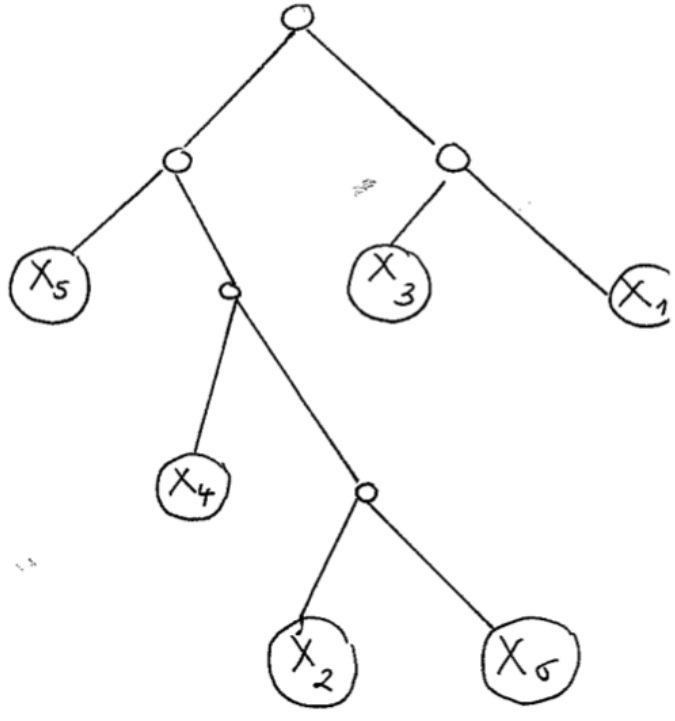
$$X_1 = bbabb, X_2 = abbaa, X_3 = babba, X_4 = ababb, X_5 = aaaab, X_6 = abbba$$

DST



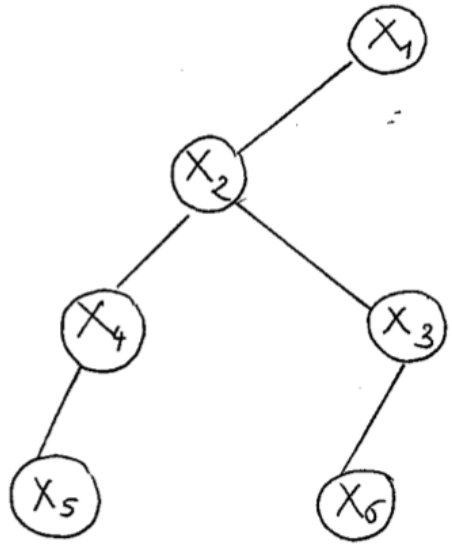
$2 + 2 \times 2 + 1 \times 3$
$= 9$

Trie



$3 \times 2 + 1 \times 3 + 2 \times 4$
$= 17$

Symbol-BST



$1 + 3 + 4 + 5 + 6$
$= 19$

General analysis of the shape of a tree structure

built on n infinite words independently emitted from the same source $\mathcal{S}$

General analysis of the shape of a tree structure

built on n infinite words independently emitted from the same source $\mathcal{S}$

A full node := a node which contains a word

- The full nodes for BST and DST : internal nodes
- The full nodes for Tries: external nodes

The depth of a node: the number of nodes between the node and the root.

Main parameters of interest

- the path length L_n := the sum of the depths of the full nodes
- the profile $B_{n,k}$:= number of full nodes at depth k
- or the typical depth D_n defined by $\Pr[D_n = k] = \frac{1}{n} B_{n,k}$.

General analysis of the shape of a tree structure

built on n infinite words independently emitted from the same source $\mathcal{S}$

A full node := a node which contains a word

- The full nodes for BST and DST : internal nodes
- The full nodes for Tries: external nodes

The depth of a node: the number of nodes between the node and the root.

Main parameters of interest

- the path length L_n := the sum of the depths of the full nodes
- the profile $B_{n,k}$:= number of full nodes at depth k
- or the typical depth D_n defined by $\Pr[D_n = k] = \frac{1}{n} B_{n,k}$.

The probabilistic behaviour of data structures built on words depends on:

- the strategy of the data structure which separates words
- the mechanism which emits words.

General analysis of the shape of a tree structure

built on n infinite words independently emitted from the same source $\mathcal{S}$

A full node := a node which contains a word

- The full nodes for BST and DST : internal nodes
- The full nodes for Tries: external nodes

The depth of a node: the number of nodes between the node and the root.

Main parameters of interest

- the path length L_n := the sum of the depths of the full nodes
- the profile $B_{n,k}$:= number of full nodes at depth k
- or the typical depth D_n defined by $\Pr[D_n = k] = \frac{1}{n} B_{n,k}$.

The probabilistic behaviour of data structures built on words depends on:

- the strategy of the data structure which separates words
- the mechanism which emits words.

For applications, importance to deal with a general source $\mathcal{S}$

Results already obtained for DST in the case of simple sources.

To be compared to Tries and BST (results obtained for general sources)

A source:= a random mechanism which emits symbols from alphabet Σ,

$$\text{The time is discrete } t = 0, t = 1, \ldots$$
$$X_i := \text{the symbol emitted at time } t = i$$

When time evolves, the source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

The source is given by the sequence $(X_0, X_1, \ldots, X_n, X_{n+1} \ldots)$

A source:= a random mechanism which emits symbols from alphabet Σ,

$$\text{The time is discrete } t = 0, t = 1, \ldots$$
$$X_i := \text{the symbol emitted at time } t = i$$

When time evolves, the source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

The source is given by the sequence $(X_0, X_1, \ldots, X_n, X_{n+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

A source:= a random mechanism which emits symbols from alphabet Σ,

$$\text{The time is discrete } t = 0, t = 1, \ldots$$
$$X_i := \text{the symbol emitted at time } t = i$$

When time evolves, the source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

The source is given by the sequence $(X_0, X_1, \ldots, X_n, X_{n+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source :

The variables X_i are independent,

with the same distribution defined by $p_i := \Pr[X_n = i]$ $(i \in \Sigma)$

A source:= a random mechanism which emits symbols from alphabet Σ,

$$\text{The time is discrete } t = 0, t = 1, \ldots$$
$$X_i := \text{the symbol emitted at time } t = i$$

When time evolves, the source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

The source is given by the sequence $(X_0, X_1, \ldots, X_n, X_{n+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source :

The variables X_i are independent,

with the same distribution defined by $p_i := \Pr[X_n = i]$ $(i \in \Sigma)$

Markov chain:

The only dependence is between consecutive X_n's

defined by the transition matrix $p_{i|j} := \Pr[X_{n+1} = i \,|X_n = j]$

A source:= a random mechanism which emits symbols from alphabet Σ,

$$\text{The time is discrete } t = 0, t = 1, \ldots$$
$$X_i := \text{the symbol emitted at time } t = i$$

When time evolves, the source produces (infinite) words of $\Sigma^{\mathbb{N}}$.
The source is given by the sequence $(X_0, X_1, \ldots, X_n, X_{n+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source :
The variables X_i are independent,
with the same distribution defined by $p_i := \Pr[X_n = i]$ $(i \in \Sigma)$

Markov chain:
The only dependence is between consecutive X_n's
defined by the transition matrix $p_{i|j} := \Pr[X_{n+1} = i \,|X_n = j]$

A general source may have many, strong correlations between its symbols.
For $w \in \Sigma^\star$, $p_w :=$ probability that a word begins with the prefix w.
The set $\{p_w, \quad w \in \Sigma^\star\}$ defines the source $\mathcal{S}$.

Analyses of Tries and Symbol–BST built on general sources

[Clément, Flajolet, V. (2001), Clément, Flajolet, Fill, V. (2009)]

Consider n words independently emitted by a general tame source $\mathcal{S}$. Then:

Analyses of Tries and Symbol–BST built on general sources

[Clément, Flajolet, V. (2001), Clément, Flajolet, Fill, V. (2009)]

Consider n words independently emitted by a general tame source $\mathcal{S}$. Then:

(i) The mean path-length of the Trie satisfies $$L_n^{[T]} \sim \frac{1}{h_{\mathcal{S}}}\, n\, \log n.$$

(ii) The mean symbol path-length of the BST satisfies $$L_n^{[B]} \sim \frac{1}{h_{\mathcal{S}}}\, n \log^2 n.$$

Analyses of Tries and Symbol–BST built on general sources

[Clément, Flajolet, V. (2001), Clément, Flajolet, Fill, V. (2009)]

Consider n words independently emitted by a general tame source $\mathcal{S}$. Then:

(i) The mean path-length of the Trie satisfies $$L_n^{[T]} \sim \frac{1}{h_{\mathcal{S}}}\, n\, \log n.$$

(ii) The mean symbol path-length of the BST satisfies $$L_n^{[B]} \sim \frac{1}{h_{\mathcal{S}}}\, n \log^2 n.$$

Here, $h_{\mathcal{S}}$ is the entropy $h_{\mathcal{S}}$ of the source $\mathcal{S}$, defined as

$$h_{\mathcal{S}} := \lim_{k\to\infty} \left[\frac{-1}{k} \sum_{w\in\Sigma^k} p_w \log p_w \right],$$

where p_w is the probability that a word begins with prefix w.

Analyses of Tries and Symbol–BST built on general sources

[Clément, Flajolet, V. (2001), Clément, Flajolet, Fill, V. (2009)]

Consider n words independently emitted by a general tame source $\mathcal{S}$. Then:

(i) The mean path-length of the Trie satisfies $$L_n^{[T]} \sim \frac{1}{h_{\mathcal{S}}}\, n\, \log n.$$

(ii) The mean symbol path-length of the BST satisfies $$L_n^{[B]} \sim \frac{1}{h_{\mathcal{S}}}\, n \log^2 n.$$

Here, $h_{\mathcal{S}}$ is the entropy $h_{\mathcal{S}}$ of the source $\mathcal{S}$, defined as

$$h_{\mathcal{S}} := \lim_{k\to\infty} \left[\frac{-1}{k} \sum_{w\in\Sigma^k} p_w \log p_w \right],$$

where p_w is the probability that a word begins with prefix w.

The remainder term depends on the "tameness" region of the source.

Previous results for a DST built on a simple source

Simple source = memoryless source or ergodic Markov chain.

To be compared to results for a Trie

Consider n words independently emitted by a simple source $\mathcal{S}$. Then:

Previous results for a DST built on a simple source

Simple source = memoryless source or ergodic Markov chain.

To be compared to results for a Trie

Consider n words independently emitted by a simple source $\mathcal{S}$. Then:

The mean internal path-length of the Digital Search Tree satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}}\, n \log n + n A_{\mathcal{S}} + R_n^{[D]}$$

Previous results for a DST built on a simple source

Simple source = memoryless source or ergodic Markov chain.

To be compared to results for a Trie

Consider n words independently emitted by a simple source $\mathcal{S}$. Then:

The mean internal path-length of the Digital Search Tree satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}}\, n\, \log n + nA_{\mathcal{S}} + R_n^{[D]}$$

The mean external path length of the Trie satisfies

$$L_n^{[T]} = \frac{1}{h_{\mathcal{S}}}\, n\, \log n + nB_{\mathcal{S}} + R_n^{[T]}$$

Previous results for a DST built on a simple source

Simple source = memoryless source or ergodic Markov chain.

To be compared to results for a Trie

Consider n words independently emitted by a simple source $\mathcal{S}$. Then:

The mean internal path-length of the Digital Search Tree satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}}\, n \log n + n A_{\mathcal{S}} + R_n^{[D]}$$

The mean external path length of the Trie satisfies

$$L_n^{[T]} = \frac{1}{h_{\mathcal{S}}}\, n \log n + n B_{\mathcal{S}} + R_n^{[T]}$$

The difference $A_{\mathcal{S}} - B_{\mathcal{S}}$ is always negative for any simple source.

The remainder terms R_n are of the same type for the two structures,

- they depend on arithmetic properties of source probabilities.
- they possibly contain a periodic term

Previous results for a DST built on a simple source

Simple source = memoryless source or ergodic Markov chain.

To be compared to results for a Trie

Consider n words independently emitted by a simple source $\mathcal{S}$. Then:

The mean internal path-length of the Digital Search Tree satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}}\, n\, \log n + nA_{\mathcal{S}} + R_n^{[D]}$$

The mean external path length of the Trie satisfies

$$L_n^{[T]} = \frac{1}{h_{\mathcal{S}}}\, n\, \log n + nB_{\mathcal{S}} + R_n^{[T]}$$

The difference $A_{\mathcal{S}} - B_{\mathcal{S}}$ is always negative for any simple source.

The remainder terms R_n are of the same type for the two structures,

- they depend on arithmetic properties of source probabilities.
- they possibly contain a periodic term

For the binary unbiased source: $\quad A_{\mathcal{S}} - B_{\mathcal{S}} = -\left[\frac{1}{\log 2} + \sum_{k\geq 1} \frac{1}{2^k - 1}\right]$

Our main result on DST's.

Our main result on DST's.

Consider a super-tame source $\mathcal{S}$. Then, the mean internal path-length of a digital search tree built on n words independently emitted by $\mathcal{S}$ satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}} n \log n + A_{\mathcal{S}}\, n + R_n^{[D]}.$$

The constant $A_{\mathcal{S}}$ is expressed with characteristics of $\mathcal{S}$

The remainder term depends on the (super)-tameness of the source,

It possibly contains a periodic term.

Our main result on DST's.

Consider a super-tame source $\mathcal{S}$. Then, the mean internal path-length of a digital search tree built on n words independently emitted by $\mathcal{S}$ satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}} n \log n + A_{\mathcal{S}}\, n + R_n^{[D]}.$$

The constant $A_{\mathcal{S}}$ is expressed with characteristics of $\mathcal{S}$

The remainder term depends on the (super)-tameness of the source,

It possibly contains a periodic term.

To be compared with the previous results obtained for tries.

Our main result on DST's.

Consider a super-tame source $\mathcal{S}$. Then, the mean internal path-length of a digital search tree built on n words independently emitted by $\mathcal{S}$ satisfies

$$L_n^{[D]} = \frac{1}{h_{\mathcal{S}}} n \log n + A_{\mathcal{S}}\, n + R_n^{[D]}.$$

The constant $A_{\mathcal{S}}$ is expressed with characteristics of $\mathcal{S}$
The remainder term depends on the (super)-tameness of the source,
It possibly contains a periodic term.

To be compared with the previous results obtained for tries.

Consider a general tame source $\mathcal{S}$. Then, the mean path-length of a trie built on n words independently emitted by $\mathcal{S}$ satisfies

$$L_n^{[T]} = \frac{1}{h_{\mathcal{S}}} n \log n + B_{\mathcal{S}}\, n + R_n^{[T]}$$

We obtain an expression for $A_{\mathcal{S}} - B_{\mathcal{S}}$ which proves that $A_{\mathcal{S}} < B_{\mathcal{S}}$

II. Sources.

A main analytical object related to any source:
the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\Lambda(1) = \infty$.

A main analytical object related to any source:

the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\quad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

A main analytical object related to any source:
the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source
– they translate them into analytic properties

For instance, the entropy $h_{\mathcal{S}}$,

$$h(\mathcal{S}) := \lim_{k \to \infty} \frac{-1}{k} \sum_{w \in \Sigma^k} p_w \log p_w = \lim_{k \to \infty} \left[-\frac{1}{k} \Lambda_k'(1) \right]$$

A main analytical object related to any source:

the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

For instance, the entropy $h_{\mathcal{S}}$,

$$h(\mathcal{S}) := \lim_{k \to \infty} \frac{-1}{k} \sum_{w \in \Sigma^k} p_w \log p_w = \lim_{k \to \infty} \left[-\frac{1}{k} \Lambda_k'(1) \right]$$

– they intervene in probabilistic analysis of algorithms and data structures.

A main analytical object related to any source:

the Dirichlet series of probabilities, $\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

A main analytical object related to any source:

the Dirichlet series of probabilities, $\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

Memoryless sources, with probabilities (p_i)

$$\Lambda(s) = \frac{1}{1 - \lambda(s)} \qquad \text{with} \quad \lambda(s) = \sum_{i=1}^{r} p_i^s$$

A main analytical object related to any source:

the Dirichlet series of probabilities, $\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

Memoryless sources, with probabilities (p_i)

$$\Lambda(s) = \frac{1}{1 - \lambda(s)} \qquad \text{with} \quad \lambda(s) = \sum_{i=1}^{r} p_i^s$$

Markov chains, defined by – the vector $\mathbf{R}$ of initial probabilities (r_i)

– and the transition matrix $\mathbf{P} := (p_{j|i})$

$$\Lambda(s) = 1 + {}^t\mathbf{R}_s (I - \mathbf{P}_s)^{-1} \mathbf{1} \qquad \text{with} \quad \mathbf{P}_s = (p_{j|i}^s), \quad \mathbf{R}_s = (r_i^s).$$

A main analytical object related to any source:

the Dirichlet series of probabilities, $\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

Memoryless sources, with probabilities (p_i)

$$\Lambda(s) = \frac{1}{1 - \lambda(s)} \qquad \text{with} \quad \lambda(s) = \sum_{i=1}^{r} p_i^s$$

Markov chains, defined by – the vector $\mathbf{R}$ of initial probabilities (r_i)

– and the transition matrix $\mathbf{P} := (p_{j|i})$

$$\Lambda(s) = 1 + {}^t\mathbf{R}_s (I - \mathbf{P}_s)^{-1} \mathbf{1} \qquad \text{with} \quad \mathbf{P}_s = (p_{j|i}^s), \quad \mathbf{R}_s = (r_i^s).$$

And for a general source?

Does $\Lambda(s)$ admit a nice alternative expression?

A general source $\mathcal{S}$ and its shifted sources.

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

A general source $\mathcal{S}$ and its shifted sources.

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

The source $\mathcal{S}$ defines a sequence of sources $\mathcal{S}_{(u)}$ (for $u \in \Sigma^\star$)

For $u \in \Sigma^\star$ with $p_u \neq 0$, the source $\mathcal{S}_{(u)} = \mathcal{S}|_u$ is a shifted source

- which gathers all the words of $\mathcal{S}$ which begin with $u \in \Sigma^\star$,
- from which the prefix u is removed.

A general source $\mathcal{S}$ and its shifted sources.

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

The source $\mathcal{S}$ defines a sequence of sources $\mathcal{S}_{(u)}$ (for $u \in \Sigma^\star$)

For $u \in \Sigma^\star$ with $p_u \neq 0$, the source $\mathcal{S}_{(u)} = \mathcal{S}|_u$ is a shifted source

- which gathers all the words of $\mathcal{S}$ which begin with $u \in \Sigma^\star$,
- from which the prefix u is removed.

The source $\mathcal{S}_{(u)}$ is completely defined

- by the fundamental (conditional) probabilities $p_{w|u}$,
- when w is any finite prefix for which $u \leq w$.

A general source $\mathcal{S}$ and its shifted sources.

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

The source $\mathcal{S}$ defines a sequence of sources $\mathcal{S}_{(u)}$ (for $u \in \Sigma^\star$)

For $u \in \Sigma^\star$ with $p_u \neq 0$, the source $\mathcal{S}_{(u)} = \mathcal{S}|_u$ is a shifted source

- which gathers all the words of $\mathcal{S}$ which begin with $u \in \Sigma^\star$,
- from which the prefix u is removed.

The source $\mathcal{S}_{(u)}$ is completely defined

- by the fundamental (conditional) probabilities $p_{w|u}$,
- when w is any finite prefix for which $u \leq w$.

In this case, w can be written as $w = u \cdot v$

The conditional probabilities $p_{w|u} = p_{(u.v)|u}$ are denoted as $q_{v|u}$.

These are the fundamental probabilities of the source $S_{(u)}$.

The generalized transition matrix of a source $\mathcal{S}$

The generalized transition matrix $\mathbf{P}$ of the source $\mathcal{S}$ ...

The generalized transition matrix of a source $\mathcal{S}$

The generalized transition matrix $\mathbf{P}$ of the source $\mathcal{S}$ ...

extends the transition matrix of a Markov chain.

This is an infinite matrix, whose rows and columns are indexed by $\Sigma^\star$

The non zero coefficients at the row u are located at the columns $u.i$

they are equal to: $$\mathbf{P}(u, u \cdot i) = \frac{p_{u \cdot i}}{p_u} = q_{i|u}$$

The generalized transition matrix of a source $\mathcal{S}$

The generalized transition matrix $\mathbf{P}$ of the source $\mathcal{S}$ ...

extends the transition matrix of a Markov chain.

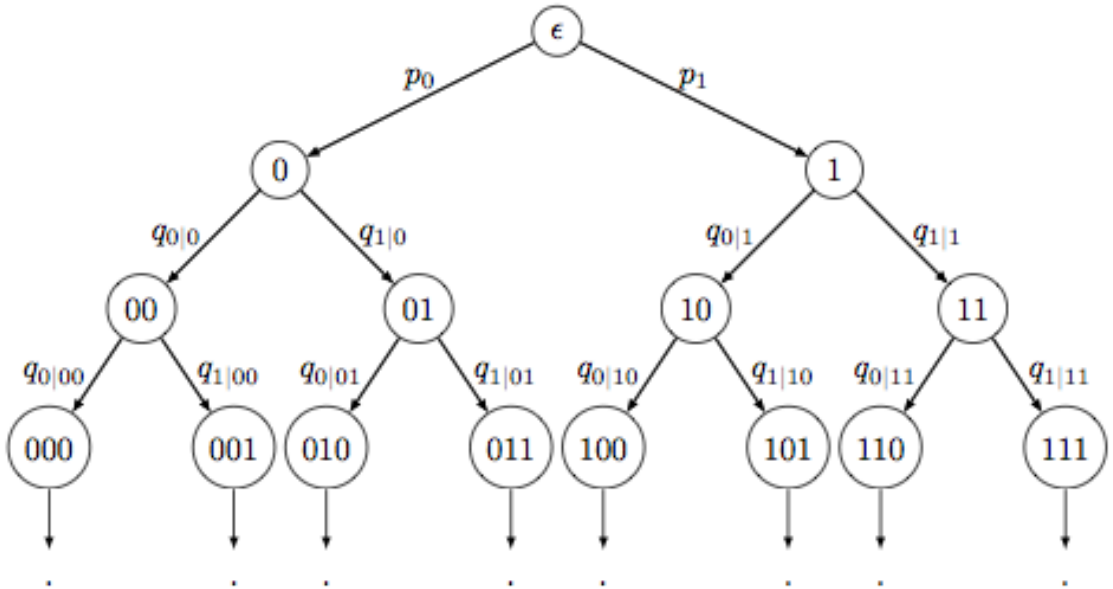
This is an infinite matrix, whose rows and columns are indexed by $\Sigma^\star$

The non zero coefficients at the row u are located at the columns $u.i$

they are equal to: $$\mathbf{P}(u, u \cdot i) = \frac{p_{u \cdot i}}{p_u} = q_{i|u}$$

The matrix $\mathbf{P}$ is the transition matrix associated to the graph of the source.

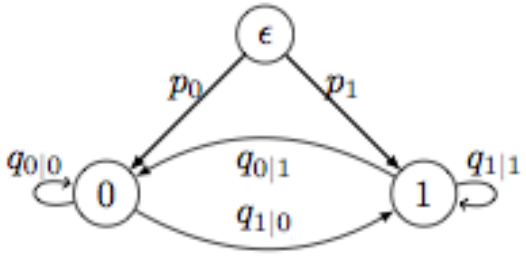
The states are the sources $\mathcal{S}_{(u)}$, the transitions are $u \to u \cdot i$.

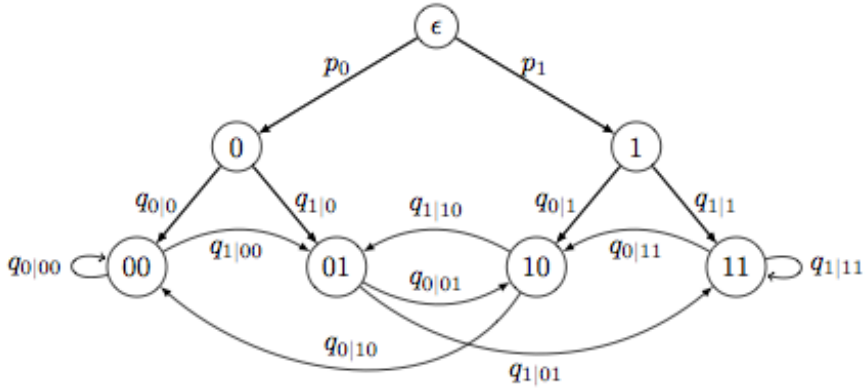


Sometimes, the graph (and thus the matrix) can be pruned:

– One only keeps the sources $S_{(u)}$ which have a different distribution

– For a simple source, this provides a finite graph (a finite matrix)

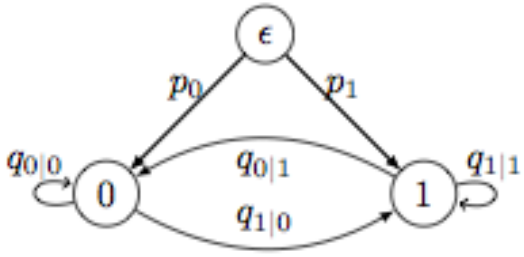


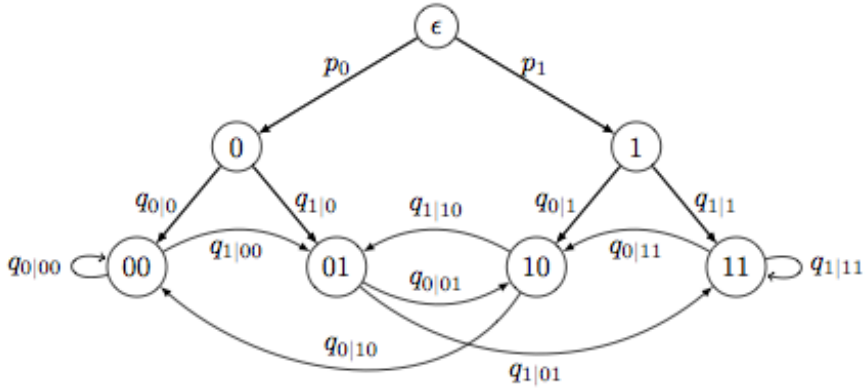


Sometimes, the graph (and thus the matrix) can be pruned:

– One only keeps the sources $S_{(u)}$ which have a different distribution

– For a simple source, this provides a finite graph (a finite matrix)





For $s \in \mathbb{C}$, the matrix $\mathbf{P}_s$ is obtained from $\mathbf{P}$

by raising its coefficients to the power s.

For $k \in \mathbb{N}$, the non zero coefficients of matrix $\mathbf{P}_s^k$ at the row u

are located at the columns $u.\alpha$ (for $\alpha \in \Sigma^k$) and equal $p^s_{u.\alpha|u} = q^s_{\alpha|u}$.

For a general source, with its generalized transition matrix $\mathbf{P}_s$, one has

$$\Lambda(s) = {}^t\mathbf{E} \cdot (I - \mathbf{P}_s)^{-1}[\mathbf{1}], \qquad \text{with} \qquad {}^t\mathbf{E} := (1, 0, 0 \ldots)$$

III. Analyses of tree data structures built on general sources.

A common strategy for the analysis of the three types of trees:
Two dictionaries –algebraic and analytic–

Source $\mathcal{S}$,		Mixed Dirichet series		The mean value L_n
its characteristics	(A)	$\varpi(s)$	(B)	of the path length
and	$\Longrightarrow$	depends both	$\Longrightarrow$	of the data structure
the data structure,		on the source		when built
its recursive definition		and the data structure		on n words of $\mathcal{S}$

A common strategy for the analysis of the three types of trees:
Two dictionaries –algebraic and analytic–

Source $\mathcal{S}$,		Mixed Dirichet series		The mean value L_n
its characteristics	(A)	$\varpi(s)$	(B)	of the path length
and	$\Longrightarrow$	depends both	$\Longrightarrow$	of the data structure
the data structure,		on the source		when built
its recursive definition		and the data structure		on n words of $\mathcal{S}$

(A) Derivation for $\varpi(s)$

First obtain a (system of) functional equations

satisfied by the Poisson generating function $B(z)$ of the path length

A common strategy for the analysis of the three types of trees:
Two dictionaries –algebraic and analytic–

Source $\mathcal{S}$,		Mixed Dirichet series		The mean value L_n
its characteristics	(A)	$\varpi(s)$	(B)	of the path length
and	$\Longrightarrow$	depends both	$\Longrightarrow$	of the data structure
the data structure,		on the source		when built
its recursive definition		and the data structure		on n words of $\mathcal{S}$

(A) Derivation for $\varpi(s)$

First obtain a (system of) functional equations
satisfied by the Poisson generating function $B(z)$ of the path length

The mixed Dirichlet series $\varpi(s)$ is related to the Mellin transform of $B(z)$.
It involves $\Lambda(s)$ or more generally the quasi-inverse $(I - \mathbf{P}_s)^{-1}$.

A common strategy for the analysis of the three types of trees:
Two dictionaries –algebraic and analytic–

Source $\mathcal{S}$,		Mixed Dirichet series		The mean value L_n
its characteristics	(A)	$\varpi(s)$	(B)	of the path length
and	$\Longrightarrow$	depends both	$\Longrightarrow$	of the data structure
the data structure,		on the source		when built
its recursive definition		and the data structure		on n words of $\mathcal{S}$

(A) Derivation for $\varpi(s)$

First obtain a (system of) functional equations
satisfied by the Poisson generating function $B(z)$ of the path length

The mixed Dirichlet series $\varpi(s)$ is related to the Mellin transform of $B(z)$.
It involves $\Lambda(s)$ or more generally the quasi-inverse $(I - \mathbf{P}_s)^{-1}$.

(B)

$$L_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi(k)$$

An exact formula, from which the asymptotics can be derived.

A common strategy for the analysis of the three types of trees:
Two dictionaries –algebraic and analytic–

Source $\mathcal{S}$, its characteristics and the data structure, its recursive definition	(A) $\Longrightarrow$	Mixed Dirichet series $\varpi(s)$ depends both on the source and the data structure	(B) $\Longrightarrow$	The mean value L_n of the path length of the data structure when built on n words of $\mathcal{S}$

(A) Derivation for $\varpi(s)$

First obtain a (system of) functional equations
satisfied by the Poisson generating function $B(z)$ of the path length

The mixed Dirichlet series $\varpi(s)$ is related to the Mellin transform of $B(z)$.
It involves $\Lambda(s)$ or more generally the quasi-inverse $(I - \mathbf{P}_s)^{-1}$.

(B) $$L_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi(k)$$

An exact formula, from which the asymptotics can be derived.
With the Rice formula, it depends on "tameness" properties of $\varpi(s)$
closely related to tameness properties of the source

Mixed Dirichlet series $\varpi(s)$ of tree structures – (I) Tries, and Symbol -BSTs

Mixed Dirichlet series $\varpi(s)$ of tree structures – (I) Tries, and Symbol -BSTs

For Tries and Symbol-BST's, the series $\varpi(s)$ is only expressed with $\Lambda(s)$:

$$\varpi_T(s) = s\Lambda(s), \qquad \varpi_B(s) = \frac{1}{s(s-1)}\Lambda(s),$$

Remind: $\Lambda(s) = {}^t\mathbf{E}\,(I - \mathbf{P}_s)^{-1}[\mathbf{1}]$, with ${}^t\mathbf{E} := (1, 0, 0\ldots)$.

Mixed Dirichlet series $\varpi(s)$ of tree structures – (I) Tries, and Symbol -BSTs

For Tries and Symbol-BST's, the series $\varpi(s)$ is only expressed with $\Lambda(s)$:

$$\varpi_T(s) = s\Lambda(s), \qquad \varpi_B(s) = \frac{1}{s(s-1)}\Lambda(s),$$

Remind: $\Lambda(s) = {}^t\mathbf{E}\,(I - \mathbf{P}_s)^{-1}[\mathbf{1}]$, with ${}^t\mathbf{E} := (1, 0, 0\ldots)$.

Tameness of the source:

Existence of a region $\mathcal{R}$ on the left of the vertical line $\Re s = 1$ where $\Lambda(s)$

- is meromorphic with a simple pole at $s = 1$
- of polynomial growth for $|\Im s| \to \infty$

Mixed Dirichlet series $\varpi(s)$ of tree structures – (I) Tries, and Symbol -BSTs

For Tries and Symbol-BST's, the series $\varpi(s)$ is only expressed with $\Lambda(s)$:

$$\varpi_T(s) = s\Lambda(s), \qquad \varpi_B(s) = \frac{1}{s(s-1)}\Lambda(s),$$

Remind: $\Lambda(s) = {}^t\mathbf{E}\,(I - \mathbf{P}_s)^{-1}[\mathbf{1}]$, with ${}^t\mathbf{E} := (1, 0, 0 \ldots)$.

Tameness of the source:

Existence of a region $\mathcal{R}$ on the left of the vertical line $\Re s = 1$ where $\Lambda(s)$

- is meromorphic with a simple pole at $s = 1$
- of polynomial growth for $|\Im s| \to \infty$

Then Rice's formula may be applied to $\varpi(s)$.

This gives the previously known results on Tries and Symbol–BST's

Mixed Dirichlet series $\varpi(s)$ of tree structures – (II) DSTs (new...)

For DST's, we prove that the series $\varpi(s)$ is also expressed with $(I-\mathbf{P}_s)^{-1}$,
more precisely with the infinite product

$$\mathbf{Q}_s := (I-\mathbf{P}_s)^{-1}(I-\mathbf{P}_{s+1})^{-1}\dots(I-\mathbf{P}_{s+k})^{-1}\dots,$$

under the form $\varpi_D(s) = {}^t\mathbf{E}\,\mathbf{Q}_s\,\mathbf{Q}_2^{-1}[\mathbf{1}] = {}^t\mathbf{E}(I-\mathbf{P}_s)^{-1}\mathbf{Q}_{s+1}\mathbf{Q}_2^{-1}[\mathbf{1}]$

Finally: $\varpi_D(s) = \Lambda(s) + A(s)$, with

$\Lambda(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{1}]$, $A(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{Q}_{s+1}-\mathbf{Q}_2][\mathbf{Q}_2^{-1}][\mathbf{1}]$

Mixed Dirichlet series $\varpi(s)$ of tree structures – (II) DSTs (new...)

For DST's, we prove that the series $\varpi(s)$ is also expressed with $(I-\mathbf{P}_s)^{-1}$,
more precisely with the infinite product

$$\mathbf{Q}_s := (I-\mathbf{P}_s)^{-1}(I-\mathbf{P}_{s+1})^{-1}\dots(I-\mathbf{P}_{s+k})^{-1}\dots,$$

under the form $\varpi_D(s) = {}^t\mathbf{E}\,\mathbf{Q}_s\,\mathbf{Q}_2^{-1}[\mathbf{1}] = {}^t\mathbf{E}(I-\mathbf{P}_s)^{-1}\mathbf{Q}_{s+1}\mathbf{Q}_2^{-1}[\mathbf{1}]$

Finally: $\varpi_D(s) = \Lambda(s) + A(s)$, with

$\Lambda(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{1}]$, $A(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{Q}_{s+1}-\mathbf{Q}_2][\mathbf{Q}_2^{-1}][\mathbf{1}]$

Super-tameness of the source:

– Existence of a region $\mathcal{R}$ on the left of the vertical line $\Re s = 1$

– Existence of a convenient functional space where

  – $s \mapsto (I-\mathbf{P}_s)^{-1}$ is meromorphic with a simple pole at $s=1$

  – where $s \mapsto ||(I-\mathbf{P}_s)^{-1}||$ is of polynomial growth for $|\Im s| \to \infty$

Mixed Dirichlet series $\varpi(s)$ of tree structures – (II) DSTs (new...)

For DST's, we prove that the series $\varpi(s)$ is also expressed with $(I-\mathbf{P}_s)^{-1}$,
more precisely with the infinite product

$$\mathbf{Q}_s := (I-\mathbf{P}_s)^{-1}(I-\mathbf{P}_{s+1})^{-1}\dots(I-\mathbf{P}_{s+k})^{-1}\dots,$$

under the form $\varpi_D(s) = {}^t\mathbf{E}\,\mathbf{Q}_s\,\mathbf{Q}_2^{-1}[\mathbf{1}] = {}^t\mathbf{E}(I-\mathbf{P}_s)^{-1}\mathbf{Q}_{s+1}\mathbf{Q}_2^{-1}[\mathbf{1}]$

Finally: $\varpi_D(s) = \Lambda(s) + A(s)$, with

$\Lambda(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{1}]$, $A(s) = {}^t\mathbf{E}\,(I-\mathbf{P}_s)^{-1}[\mathbf{Q}_{s+1}-\mathbf{Q}_2][\mathbf{Q}_2^{-1}][\mathbf{1}]$

Super-tameness of the source:

– Existence of a region $\mathcal{R}$ on the left of the vertical line $\Re s = 1$

– Existence of a convenient functional space where

  – $s \mapsto (I-\mathbf{P}_s)^{-1}$ is meromorphic with a simple pole at $s=1$

  – where $s \mapsto ||(I-\mathbf{P}_s)^{-1}||$ is of polynomial growth for $|\Im s| \to \infty$

Then Rice's formula may be applied to $\varpi(s)$ and the theorem is proven.

IV. Some steps of the proof.

We consider a DST built on n words independently emitted by a source $\mathcal{S}$.
We study the random variable path length $\ell_n := \ell_n(\mathcal{S})$, and deal with all the sources $\mathcal{S}_{(w)}$ and all the random variables $\ell_n^{(w)} := \ell_n(\mathcal{S}_{(w)})$

We consider a DST built on n words independently emitted by a source $\mathcal{S}$.

We study the random variable path length $\ell_n := \ell_n(\mathcal{S})$, and deal with all the sources $\mathcal{S}_{(w)}$ and all the random variables $\ell_n^{(w)} := \ell_n(\mathcal{S}_{(w)})$

$\ell_n^{(w)}$ = the path length of a DST of size n built on $\mathcal{S}_{(w)}$,

depends on two indices:

- the cardinality $n \in \mathbb{N}$
- the prefix $w \in \Sigma^\star$ which defines the source $\mathcal{S}_{(w)}$.

We consider a DST built on n words independently emitted by a source $\mathcal{S}$.

We study the random variable path length $\ell_n := \ell_n(\mathcal{S})$, and deal with all the sources $\mathcal{S}_{(w)}$ and all the random variables $\ell_n^{(w)} := \ell_n(\mathcal{S}_{(w)})$

$\ell_n^{(w)}$ = the path length of a DST of size n built on $\mathcal{S}_{(w)}$,

depends on two indices:

– the cardinality $n \in \mathbb{N}$

– the prefix $w \in \Sigma^\star$ which defines the source $\mathcal{S}_{(w)}$.

For $\Sigma := \{0, 1\}$, the basic recurrence for the sequence $\ell_n^{(w)}$ is

$$\ell_n{}^{(w)} = n - 1 + \ell_{K_n}{}^{(w.0)} + P_{n-1-K_n}{}^{(w.1)}$$

The number of nodes $K_n := K_n^{(w)}$ in the left subtree

follows a binomial law of parameter $q_{0|w}$.

$$\Pr[K_n{}^{(w)} = k] = \binom{n-1}{k} q_{0|w}^k \, q_{1|w}^{n-1-k}$$

We consider a DST built on n words independently emitted by a source $\mathcal{S}$.
We study the random variable path length $\ell_n := \ell_n(\mathcal{S})$, and deal with all the sources $\mathcal{S}_{(w)}$ and all the random variables $\ell_n^{(w)} := \ell_n(\mathcal{S}_{(w)})$

$\ell_n^{(w)}$ = the path length of a DST of size n built on $\mathcal{S}_{(w)}$,

depends on two indices:

- the cardinality $n \in \mathbb{N}$
- the prefix $w \in \Sigma^\star$ which defines the source $\mathcal{S}_{(w)}$.

For $\Sigma := \{0, 1\}$, the basic recurrence for the sequence $\ell_n^{(w)}$ is

$$\ell_n{}^{(w)} = n - 1 + \ell_{K_n}{}^{(w.0)} + P_{n-1-K_n}{}^{(w.1)}$$

The number of nodes $K_n := K_n^{(w)}$ in the left subtree

follows a binomial law of parameter $q_{0|w}$.

$$\Pr[K_n{}^{(w)} = k] = \binom{n-1}{k} q_{0|w}^k \, q_{1|w}^{n-1-k}$$

Then, the basic recurrence for the expectations $L_n{}^{(w)} := \mathbb{E}[\ell_n^{(w)}]$ is

$$L_n{}^{(w)} = n - 1 + \sum_{k=0}^{n-1} \binom{n-1}{k} q_{0|w}^k \, q_{1|w}^{n-1-k} \left(L_k^{(w\cdot 0)} + L_{n-1-k}^{(w\cdot 1)} \right)$$

$$L_n{}^{(w)} = n - 1 + \sum_{k=0}^{n-1} \binom{n-1}{k} q_{0|w}^{k} \, q_{1|w}^{n-1-k} \left(L_{k}^{(w\cdot 0)} + L_{n-1-k}^{(w\cdot 1)} \right)$$

$$L_n{}^{(w)} = n-1 + \sum_{k=0}^{n-1} \binom{n-1}{k} q_{0|w}^{k}\, q_{1|w}^{n-1-k} \left(L_k^{(w\cdot 0)} + L_{n-1-k}^{(w\cdot 1)} \right)$$

The Poisson generating functions $B^{(w)}(z) := e^{-z} \sum_{n\geq 0} L_n^{(w)} \frac{z^n}{n!}$ satisfy

$$\frac{d}{dz} B^{(w)}(z) + B^{(w)}(z) = z + B^{(w.0)}(q_{0|w} z) + B^{(w.1)}(q_{1|w} z).$$

$$L_n{}^{(w)} = n-1 + \sum_{k=0}^{n-1} \binom{n-1}{k} q_{0|w}^{k}\, q_{1|w}^{n-1-k} \left(L_k^{(w\cdot 0)} + L_{n-1-k}^{(w\cdot 1)} \right)$$

The Poisson generating functions $B^{(w)}(z) := e^{-z} \sum_{n\geq 0} L_n^{(w)} \frac{z^n}{n!}$ satisfy

$$\frac{d}{dz} B^{(w)}(z) + B^{(w)}(z) = z + B^{(w.0)}(q_{0|w} z) + B^{(w.1)}(q_{1|w} z).$$

Easy extension to any alphabet Σ:

$$\frac{d}{dz} B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w} z).$$

$$L_n{}^{(w)} = n-1 + \sum_{k=0}^{n-1} \binom{n-1}{k} q_{0|w}^{k}\, q_{1|w}^{n-1-k} \left(L_k^{(w\cdot 0)} + L_{n-1-k}^{(w\cdot 1)} \right)$$

The Poisson generating functions $B^{(w)}(z) := e^{-z} \sum_{n\geq 0} L_n^{(w)} \frac{z^n}{n!}$ satisfy

$$\frac{d}{dz} B^{(w)}(z) + B^{(w)}(z) = z + B^{(w.0)}(q_{0|w} z) + B^{(w.1)}(q_{1|w} z).$$

Easy extension to any alphabet Σ:

$$\frac{d}{dz} B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w} z).$$

This system of functional equations involves both

- the derivation d/dz
- the change of variables $z \mapsto qz$
- the shift on words $w \mapsto w.i$

In comparison, for tries, the derivation does not occur.

Different possible steps of the analysis

Basic equations $$\frac{d}{dz}B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z).$$

Different possible steps of the analysis

Basic equations $\quad \frac{d}{dz}B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z).$

Laplace $\Downarrow$

Exact expression of $B(z)$ $\quad \Longrightarrow$ AlgDePo $\Longrightarrow$ Exact expression of L_n

Different possible steps of the analysis

Basic equations $\quad \dfrac{d}{dz}B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z).$

Laplace $\Downarrow$

Exact expression of $B(z)$ $\Longrightarrow$ AlgDePo $\Longrightarrow$ Exact expression of L_n

$$B(z) = \sum_{w\in\Sigma^\star} \delta(v)[e^{-zp_v} - 1 + zp_v] \qquad L_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_D(k)$$

$$\delta(v) = \frac{1}{p_v} \sum_{w\geq v} p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1 - p_v p_\alpha^{-1}} \qquad \varpi_D(s) = \sum_{w\in\Sigma^\star} \delta(v) p_v^s,$$

Different possible steps of the analysis

Basic equations $\quad \dfrac{d}{dz}B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z).$

Laplace $\Downarrow$

Exact expression of $B(z)$ $\Longrightarrow$ AlgDePo $\Longrightarrow$ Exact expression of L_n

$$B(z) = \sum_{w\in\Sigma^\star} \delta(v)[e^{-zp_v} - 1 + zp_v] \qquad L_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_D(k)$$

$$\delta(v) = \frac{1}{p_v} \sum_{w\geq v} p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1 - p_v p_\alpha^{-1}} \qquad \varpi_D(s) = \sum_{w\in\Sigma^\star} \delta(v) p_v^s,$$

Mellin $\Downarrow$ $\qquad$ $\Downarrow$ Rice

Different possible steps of the analysis

Basic equations $\quad \dfrac{d}{dz}B^{(w)}(z) + B^{(w)}(z) = z + \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z).$

Laplace $\Downarrow$

Exact expression of $B(z)$ $\Longrightarrow$ AlgDePo $\Longrightarrow$ Exact expression of L_n

$$B(z) = \sum_{w\in\Sigma^\star} \delta(v)[e^{-zp_v} - 1 + zp_v] \qquad L_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_D(k)$$

$$\delta(v) = \frac{1}{p_v} \sum_{w\geq v} p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1 - p_v p_\alpha^{-1}} \qquad \varpi_D(s) = \sum_{w\in\Sigma^\star} \delta(v) p_v^s,$$

Mellin $\Downarrow$ $\qquad$ Rice $\Downarrow$

Alternative expression of $\varpi(s)$ $\qquad$ Asymptotics of L_n

$\varpi_D(s) = {}^t\mathbf{E}\,\mathbf{Q}_s\,\mathbf{Q}_2^{-1}[\mathbf{1}]$ $\qquad$ with tameness of $(I - \mathbf{P}_s)^{-1}$

Various extensions

Various extensions

– Already studied : the distribution of the typical depth.

Main result. For a source assumed to be hyper-tame and log–convex, the typical depth of a DST follows an asymptotic gaussian law.

The asymptotic mean values of the mean and the variance involve the same constants as in the case of tries.

Various extensions

– Already studied : the distribution of the typical depth.

Main result. For a source assumed to be hyper-tame and log–convex, the typical depth of a DST follows an asymptotic gaussian law.
The asymptotic mean values of the mean and the variance involve the same constants as in the case of tries.

– To be done :

  – Return to the analysis of the Lempel Ziv algorithm.

Various extensions

– Already studied : the distribution of the typical depth.

Main result. For a source assumed to be hyper-tame and log–convex, the typical depth of a DST follows an asymptotic gaussian law.
The asymptotic mean values of the mean and the variance involve the same constants as in the case of tries.

– To be done :

  – Return to the analysis of the Lempel Ziv algorithm.

  – Make precise all the tameness properties,
    (tame, super-tame, hyper-tame)
    even in the case of simple sources
    Work in progress with E. Cesaratto, J. Clément.