Digital search trees

Analysis of different digital trees with Rice’s integrals.

JASS

Nicolai v. Hoyningen-Huene

28.3.2004

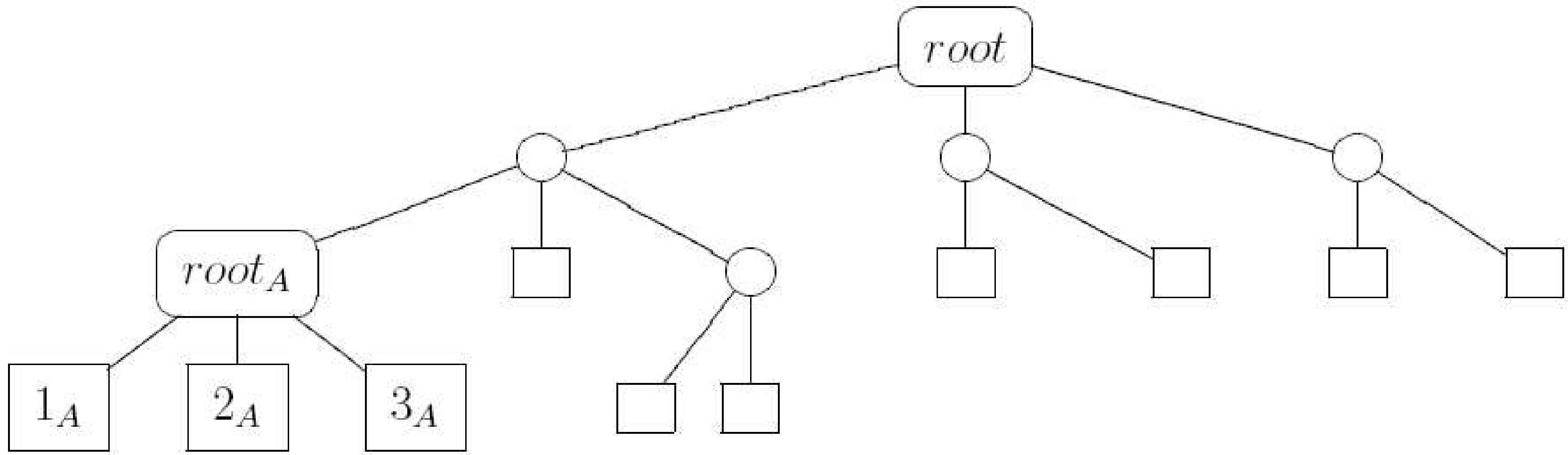
content

- **Tree**
- **Digital search tree:**
  - Definition
  - Average case analysis
- **Tries:**
  - Definition
  - Average case analysis
- **General framework**

content

- Tree
- Digital search tree
- Tries
- General framework

Tree



Tree

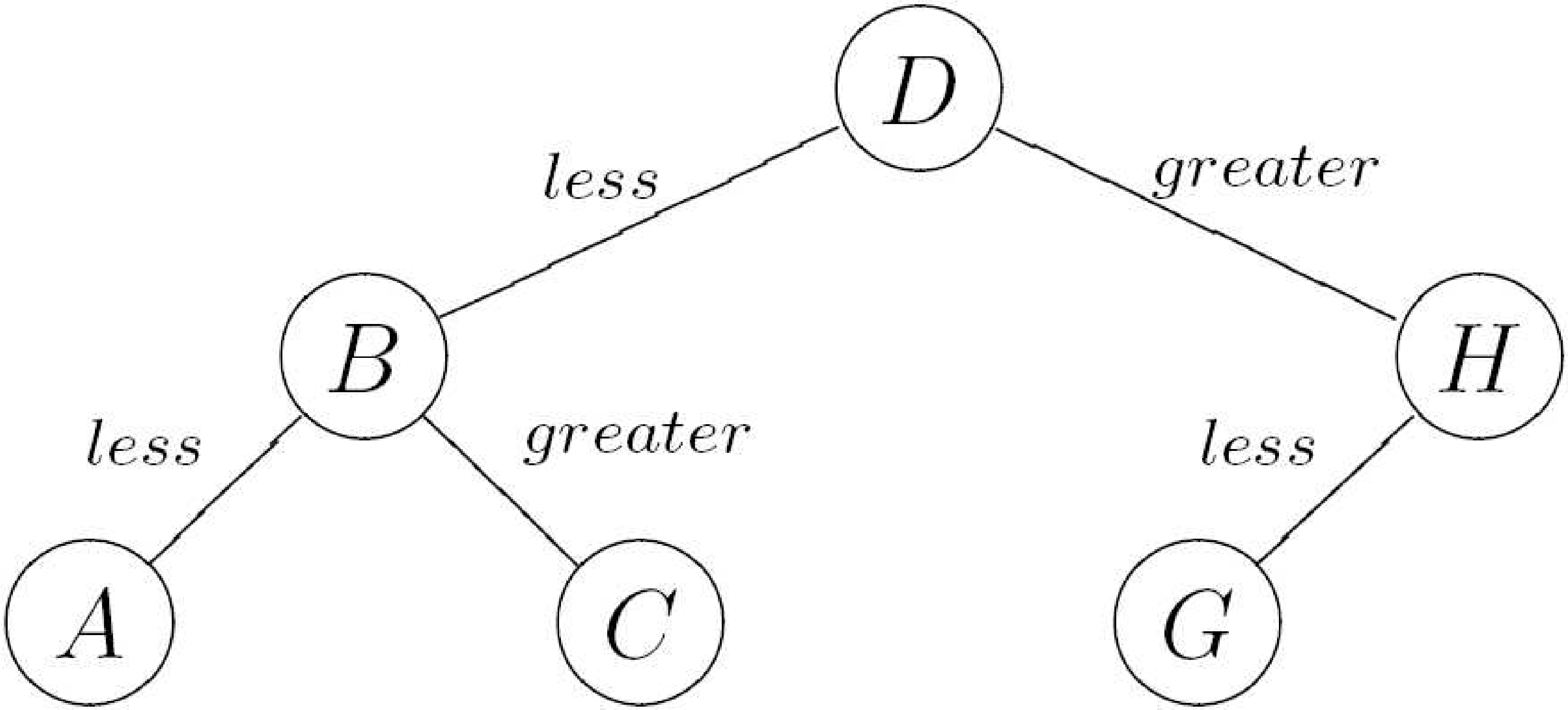
Definition 0.1 *A **tree** is defined in several ways:*

1. *A connected, undirected, acyclic graph. It is rooted and ordered unless otherwise specified.*

2. *A data structure accessed beginning at the root node. Each node is either a leaf or an internal node. An internal node has one or more child nodes and is called the parent of its child nodes. All children of the same node are siblings.*

3. *A tree is either empty (no nodes), or a root and zero or more subtrees. The subtrees are ordered.*

Search tree



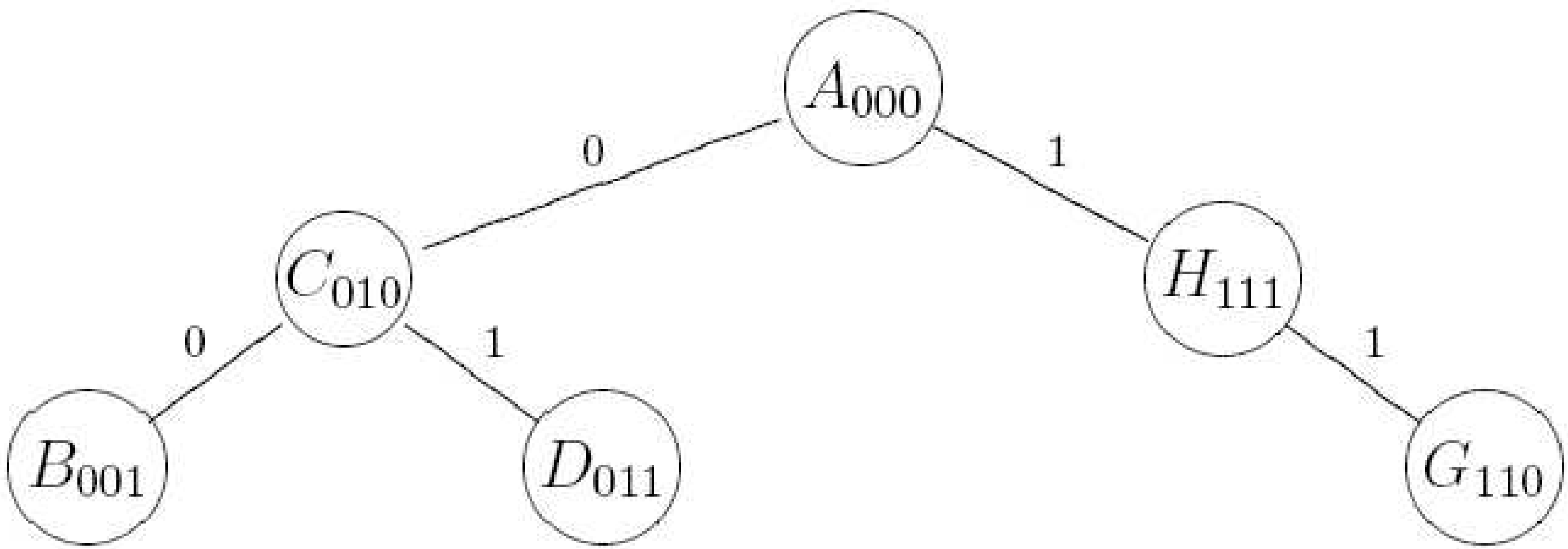
content

- **Tree**
- **Digital search tree:**
  - Definition
  - Average case analysis
- **Tries**
- **General framework**

content

- **Tree**
- **Digital search tree:**
- **Definition**
  - Average case analysis
- **Tries**
- **General framework**

Digital search tree



Digital search tree

A **digital search tree** is a dictionary implemented as a digital tree which stores strings in internal nodes, so there is no need for extra leaf nodes to store the strings.

content

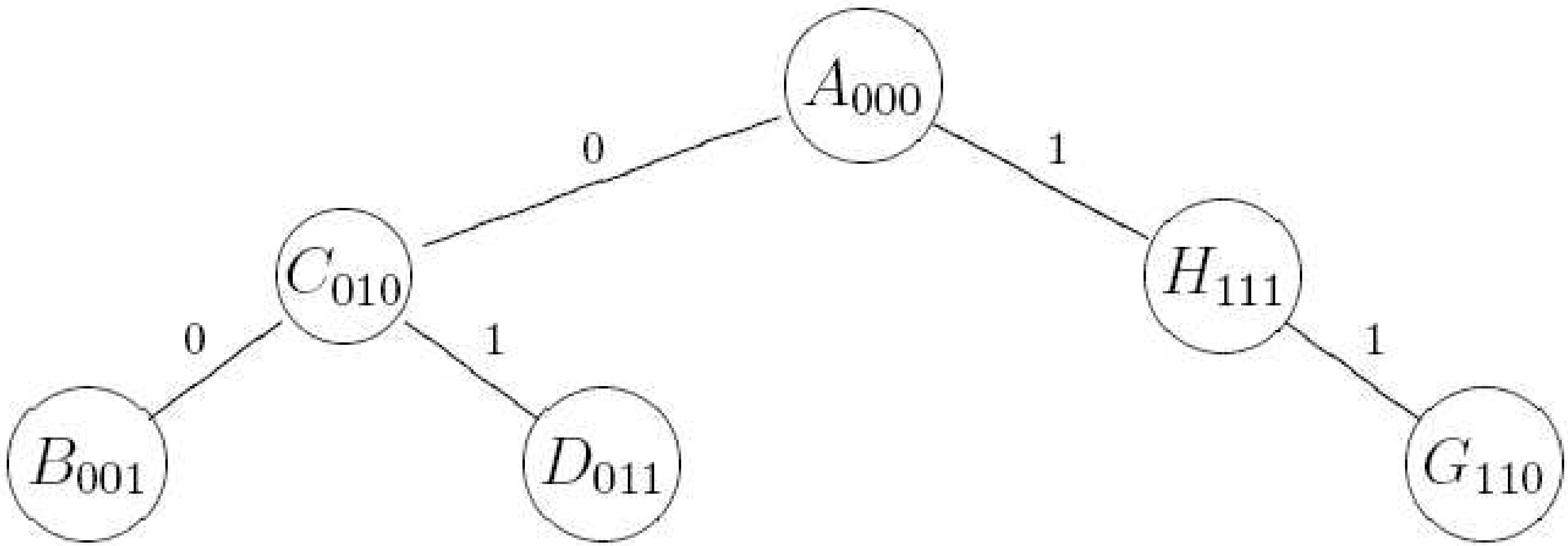
- **Tree**
- **Digital search tree:**
  - Definition
- **Average case analysis:**
  - – Internal path length
  - – External internal nodes
- **Tries**
- **General framework**

content

- **Tree**
- **Digital search tree:**
  - Definition
  - Average case analysis:
- **Internal path length**
  - – External internal nodes
- **Tries**
- **General framework**

Internal path length

The internal path length of a tree is the sum of the depth of every node of the tree.



Internal path length

Fundamental recurrence relation

$$A_N = N-1+\sum_{k=0}^{\infty} \frac{1}{2^{N-1}} \binom{N-1}{k} \left(A_k + A_{N-1-k}\right), \qquad N \geq 1$$

with $A_0 := 0$.

Internal path length

Transformation

$$\sum_{N=1}^{\infty} \frac{A_N z^{N-1}}{(N-1)!} = ze^z + 2\sum_{k=0}^{\infty} \frac{A_k}{k!} \left(\frac{z}{2}\right)^k e^{\frac{z}{2}}$$

$$A'(z) = ze^z + 2A\left(\frac{z}{2}\right) e^{\frac{z}{2}}$$

Internal path length

Substitution by $B(z)$

$$A\left(z\right)=e^{z}B\left(z\right)=\left(\sum_{N=0}^{\infty}\frac{z^{N}}{N!}\right)\left(\sum_{N=0}^{\infty}B_{N}\frac{z^{N}}{N!}\right)$$

$$A_{N}=\sum_{k=0}^{N}\binom{N}{k}B_{k}$$

Internal path length

Substitution by $B(z)$

$$A'(z) = ze^z + 2A\left(\frac{z}{2}\right)e^{\frac{z}{2}}$$

$$B'(z) + B(z) = z + 2B\left(\frac{z}{2}\right)$$

$$B_N + B_{N-1} = \frac{1}{2^{N-2}}B_{N-1}$$

Internal path length

Substitution by $B(z)$

$$B_N = (-1)^N \prod_{j=1}^{N-2} \left(1 - \frac{1}{2^j}\right)$$

Internal path length

Introduction of Q_N

$$Q_N = \prod_{j=1}^{N} \left(1 - \frac{1}{2^j}\right)$$

Internal path length

Introduction of $Q(x)$

$$Q(x) = \prod_{j=1}^{\infty} \left(1 - \frac{x}{2^j}\right)$$

$$Q(1) = Q_{\infty}$$

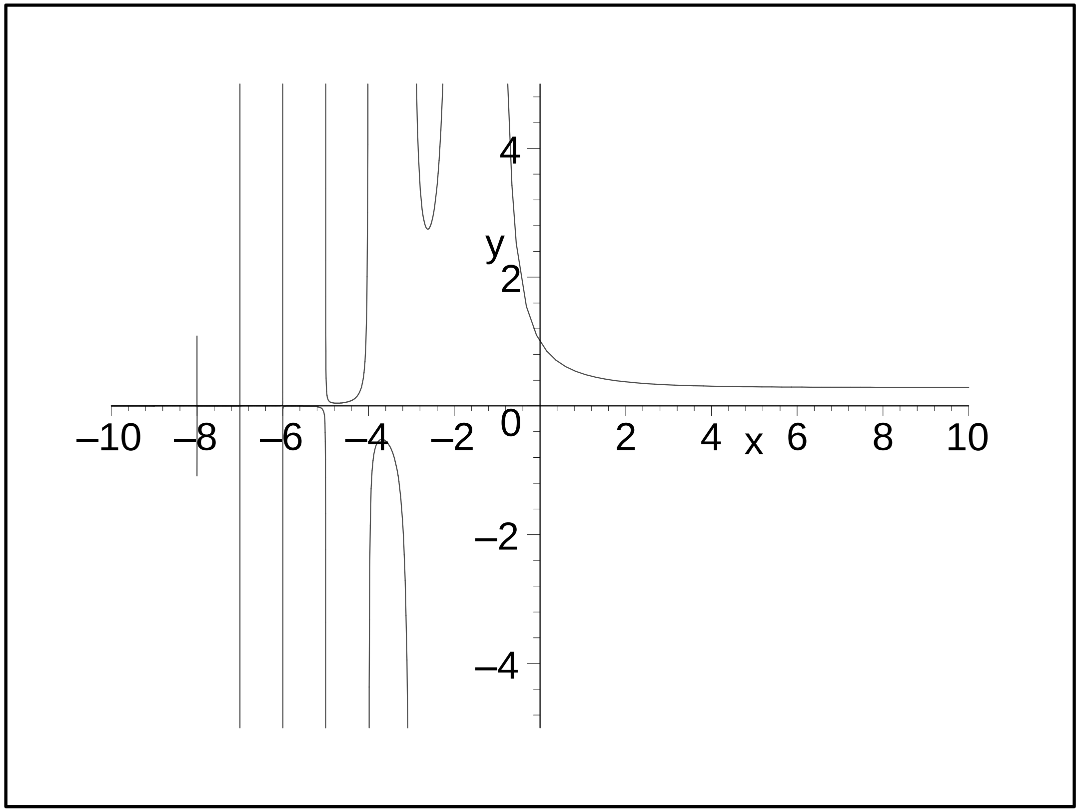
Internal path length

$Q(x)$ is used for a meromorphic function of Q_N

$$Q_N = \frac{Q\left(1\right)}{Q\left(2^{-N}\right)}$$

Internal path length

The function $Q(x)$



Internal path length

Explicit formula for A_N

$$A_N = \sum_{k=2}^{N} \binom{N}{k} (-1)^k \frac{Q(1)}{Q(2^{-k+2})}$$

Internal path length

Rice's method

$$\sum_{k=0}^{N} \binom{N}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_C B(N+1, -z) f(z)\, dz$$

Internal path length

Application of Rice's method

$$A_N = -\frac{1}{2\pi i}\int_C B\left(N+1,-z\right)\frac{Q\left(1\right)}{Q\left(2^{-z+2}\right)}dz$$

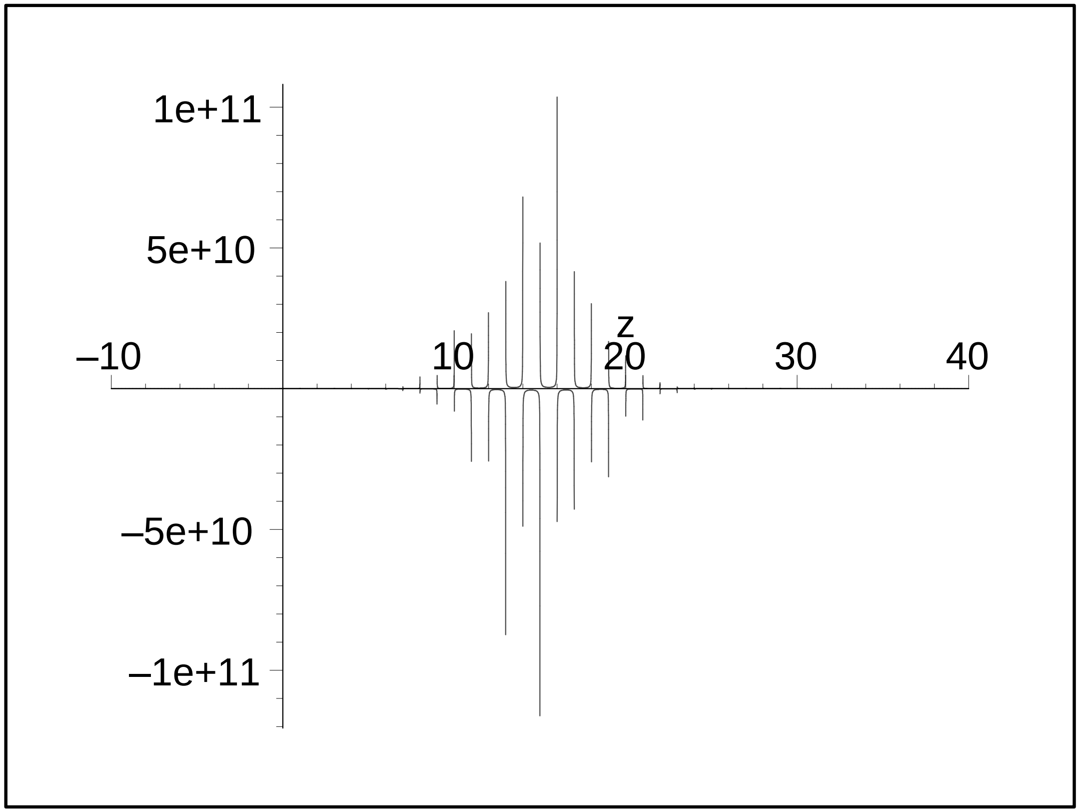
Internal path length

Beta-Function

$$B\left(p,q\right)=\frac{\Gamma\left(p\right)\Gamma\left(q\right)}{\Gamma\left(p+q\right)}$$

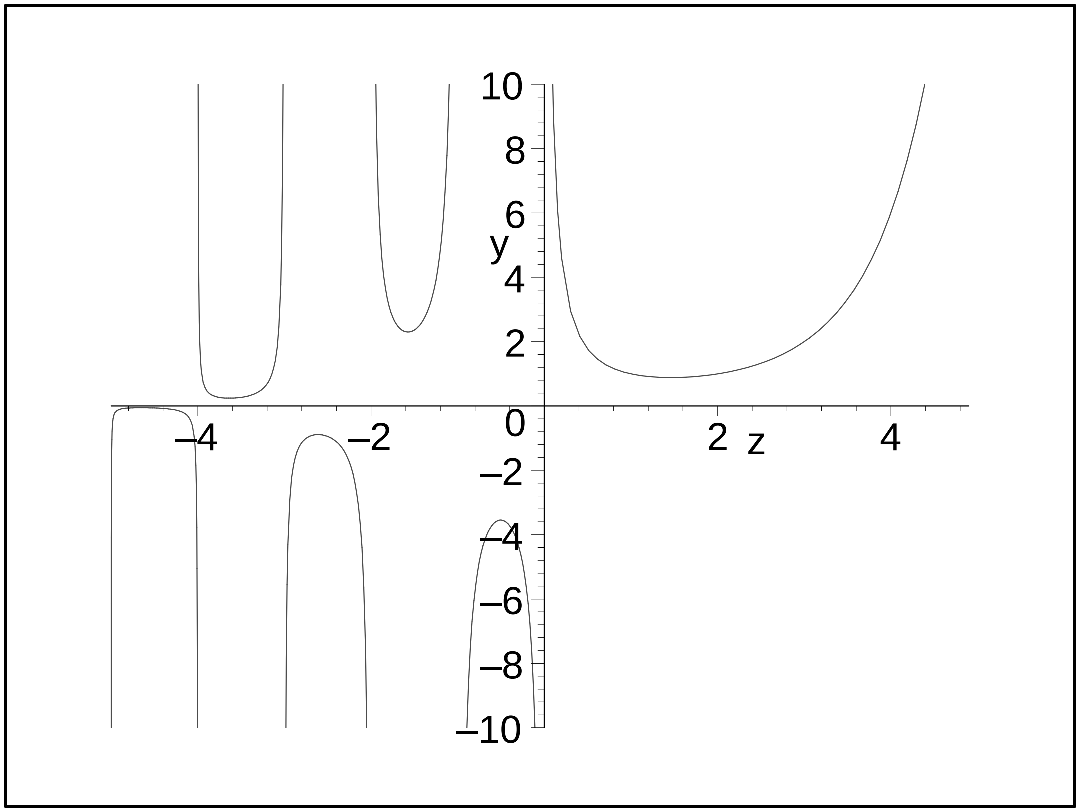
Internal path length

Beta-Function



Internal path length

Gamma-Function



Internal path length

Cauchy's theorem

If $f(z)$ is analytic in C except final number of poles $a_1, a_2, ..., a_n$ inside C, then

$$\frac{1}{2\pi i}\int_C f\left(z\right)dz = \sum_{k=1}^{n} Res_{z=a_k} f\left(z\right)$$

Internal path length

Approximation with Rectangle R_{XY} **with** $(\frac{1}{2} \pm iY, X \pm iY)$

$$A_N = -\frac{1}{2\pi i} \int_{R_{XY}} B\left(N+1, -z\right) \frac{Q\left(1\right)}{Q\left(2^{-z+2}\right)} dz - Res_{R_{XY}/C}$$

Internal path length

Approximation of the integral

$$\frac{1}{2\pi i}\int_{R_{XY}} B\left(N+1,-z\right)\frac{Q\left(1\right)}{Q\left(2^{-z+2}\right)}dz =$$

$$O\left(\int_{-Y}^{Y}\frac{\Gamma\left(N+1\right)}{\Gamma\left(N+\frac{1}{2}-iy\right)}dy\right) = O\left(\int_{-Y}^{Y} N^{\frac{1}{2}-iy}dy\right) = O\left(N^{\frac{1}{2}}\right)$$

Internal path length

Residue at $z = 1$

$$-B\left(N+1,-z\right) = -\frac{N}{z-1} - N\left(H_{N-1} - 1\right) + O\left(z-1\right)$$

$$H_{N-1} = \gamma + \ln N - O\left(\frac{1}{N}\right)$$

$$-B\left(N+1,-z\right) = -\frac{N}{z-1} - N\left(\gamma + \ln N - 1\right) + O\left(z-1\right)$$

Internal path length

Residue at $z = 1$

$$\frac{Q(1)}{Q\left(2^{-z+1}\right)} = 1 - \alpha \ln 2 \,(z-1) + O\left((z-1)^2\right)$$

Internal path length

Residue at $z = 1$

$$
\begin{aligned}
\frac{1}{1-2^{-z+1}} &= \frac{1}{1-e^{\ln 2(-z+1)}} \\
&= -\frac{1}{(-z+1)\ln 2} + \frac{1}{2} - \frac{-z+1}{12} + O\left((-z+1)^3\right) \\
&= \frac{1}{(z-1)\ln 2} + \frac{1}{2} + O\left(z-1\right)
\end{aligned}
$$

Internal path length

Residue at $z = 1$

$$\Delta_{z=1} = -N \lg N - N\left(\frac{\gamma - 1}{\ln 2} - \alpha + \frac{1}{2}\right) + O\left(1\right)$$

Internal path length

Residues at $z = j \pm \frac{2\pi ik}{\ln 2}$ **for** $Q(2^{-z+j})$

$$\Delta_{z=1\pm\frac{2\pi Ik}{\ln 2}} = -N\delta\left(N\right) + O\left(1\right)$$

where

$$\delta\left(N\right) = \frac{1}{\ln 2}\sum_{k\neq 0}\Gamma\left(-1 - \frac{2\pi ik}{\ln 2}\right)e^{2\pi ik\lg N}$$

Internal path length

Average case

$$A_N = N \lg N + N \left(\frac{\gamma - 1}{\ln 2} - \alpha + \frac{1}{2} + \delta(N) \right) + O\left(N^{\frac{1}{2}} \right)$$

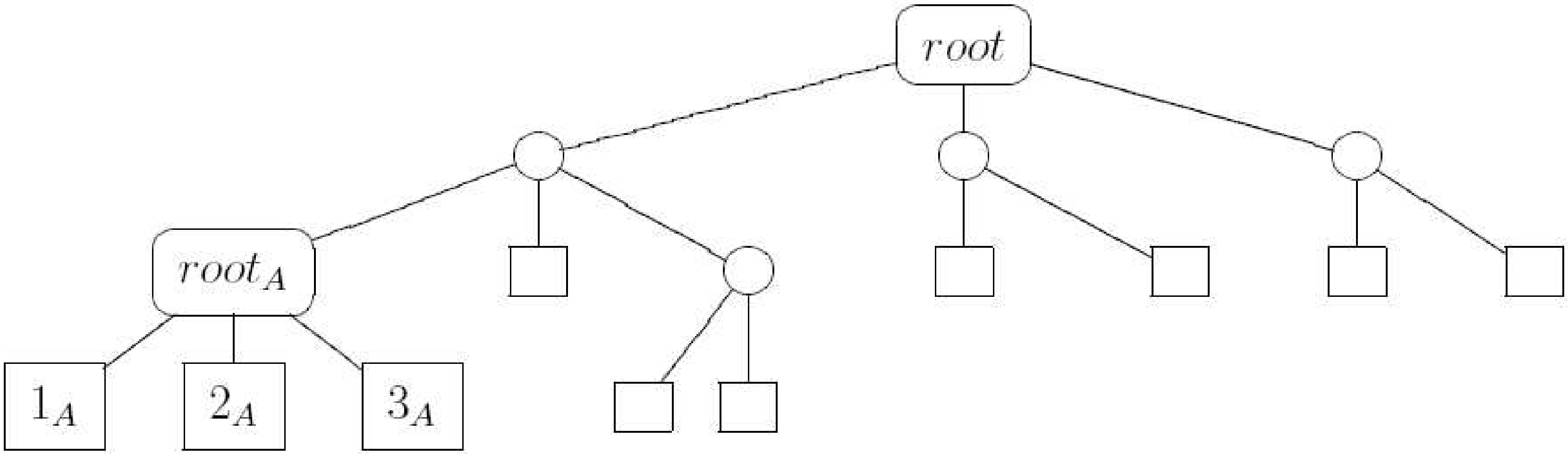
content

- **Tree**
- **Digital search tree:**
  - Definition
  - Average case analysis:
    - – Internal path length
    - **External internal nodes**
    - – Multiway branching
- **Tries**
- **General framework**

External internal nodes

Definition

External internal nodes are nodes with both links null.



External internal nodes

Fundamental recurrence relation

$$C_N = \sum_{k=0}^{\infty} \frac{1}{2^{N-1}} \binom{N-1}{k} \left(C_k + C_{N-1-k}\right), \qquad N \geq 2$$

with $C_1 = 1$ and $C_0 = 0$.

External internal nodes

Transformation

$$C(z) = \sum_{N=0}^{\infty} \frac{C_N z^N}{N!}$$

$$C'(z) = 1 + 2C\left(\frac{z}{2}\right) e^{\frac{z}{2}}$$

External internal nodes

Substitution by $D\left(z\right) = e^{-z} C\left(z\right)$

$$D\left(z\right) = \sum_{N=0}^{\infty} \frac{D_N z^N}{N!}$$

$$D'\left(z\right) + D\left(z\right) = e^{-z} + 2D\left(\frac{z}{2}\right)$$

External internal nodes

Recurrence for D_N

$$D_N + D_{N-1} = (-1)^{N-1} + \frac{1}{2^{N-2}} D_{N-1}$$

$$D_N = (-1)^{N-1} - \left(1 - \frac{1}{2^{N-2}}\right) D_{N-1}, \qquad N \geq 2$$

with $D_1 = 1$ and $D_0 = 0$

External internal nodes

Introduction of R_N

$$R_N = Q_N \left(1 + \sum_{k=1}^{N} \frac{1}{Q_k} \right)$$

External internal nodes

Explicit formula for C_N

$$C_N = N - \sum_{k=2}^{\infty} \binom{N}{k} (-1)^k R_{k-2}$$

External internal nodes

Simpler coefficients R_N^*

$$R_N^* = \frac{(N+1-\alpha)\,q^{N+1}}{1-q^{N+1}} + \frac{1}{1-q^{N+1}} R_{N+1}^*$$

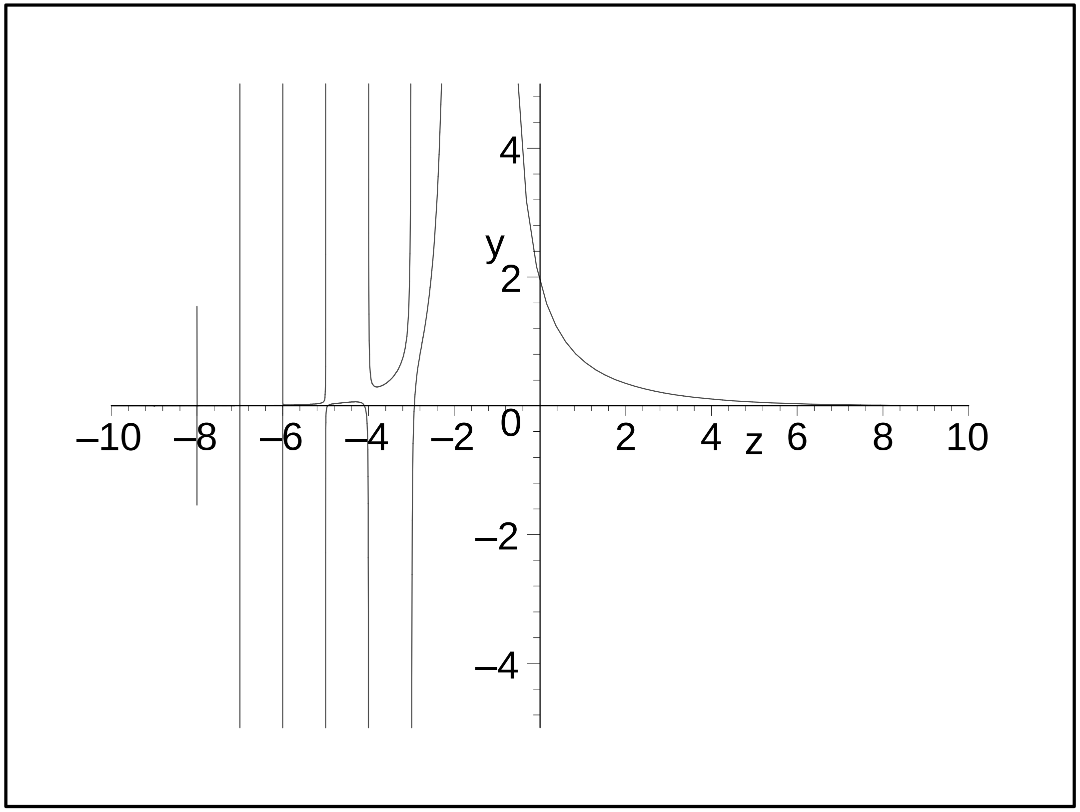
External internal nodes

The meromorphic function $R^*(z)$

$$R^*(z) = \sum_{i=2}^{\infty} \frac{(z+1+i-\alpha)\, q^{z+1+i}}{\prod_{j=0}^{i} (1-q^{z+1+j})}$$

External internal nodes

The meromorphic function $R^*(z)$



External internal nodes

Explicit formula for C_N

$$C_N = (N-1)(\alpha+1) - \sum_{k=2}^{\infty} \binom{N}{k} (-1)^k R^*_{k-2}$$

External internal nodes

Applying Rice's method

$$C_N - (N-1)(\alpha+1) = \frac{1}{2\pi i}\int_C B(N+1,-z)\,R^*(z-2)\,dz$$

External internal nodes

Approximation with Rectangle R_{XY}

$$C_N - (N-1)(\alpha+1) = \frac{1}{2\pi i}\int_{R_{XY}} B(N+1,-z)\,R^*(z-2)\,dz - \Delta$$

External internal nodes

Approximation of the integral

$$O\left(\int_{-Y}^{Y} \frac{\Gamma\left(N+1\right)}{\Gamma\left(N+\frac{1}{2}-iy\right)}dy\right) = O\left(\int_{-Y}^{Y} N^{\frac{1}{2}-iy}dy\right) = O\left(N^{\frac{1}{2}}\right)$$

External internal nodes

Residue at $z = 1$

$$\Delta_{z=1} = N\left(\beta + 1 - \frac{1}{Q_\infty}\left(\alpha^2 - \alpha - \frac{1}{\ln q}\right)\right)$$

External internal nodes

Residues at $z = -1 \pm \frac{2\pi i k}{\ln q}$

$$\delta^*(N) = \frac{2\pi i k}{Q_\infty \ln q} \sum_{k \neq 0} \frac{1}{\ln q} \Gamma\left(-1 - \frac{2\pi i k}{\ln q}\right) e^{2\pi i k \lg N}$$

External internal nodes

Average case

$$C_N = N\left(\beta + 1 - \frac{1}{Q_\infty}\left(\frac{1}{\ln 2} + \alpha^2 - \alpha\right) + \delta^*\left(N\right)\right) + O\left(N^{\frac{1}{2}}\right)$$

content

- **Tree**
- **Digital search tree:**
  - Definition
  - Average case analysis:
    - Internal path length
    - External internal nodes
- **Multiway branching**
- **Tries**
- **General framework**

Multiway branching

Fundamental recurrence relation for external nodes

$$C_N^{[M]} = \sum_{k_1+k_2+...+k_M=N-1} \frac{1}{M^{N-1}} \binom{N-1}{k_1, k_2, ..., k_M} \left(\sum_{i=1}^{M} C_{k_i}^{[M]} \right)$$

with $C_1^{[M]} = 1$ and $C_0^{[M]} = 0$

Multiway branching

Fundamental recurrence relation for external nodes

$$C_N^{[M]} = M \sum_{k_1+k_2+...+k_M=N-1} \frac{1}{M^{N-1}} \binom{N-1}{k_1, k_2, ..., k_M} C_{k_1}^{[M]}$$

with $C_1^{[M]} = 1$ and $C_0^{[M]} = 0$

Multiway branching

Transformation

$$C^{[M]}(z) = \sum_{N=0}^{\infty} \frac{C_N^{[M]} z^N}{N!}$$

$$C^{[M]\prime}(z) = 1 + MC^{[M]}\left(\frac{z}{M}\right)\left(e^{\left(1-\frac{1}{M}\right)z}\right)$$

Multiway branching

average case for external nodes

$$
\begin{aligned}
C_N^{[M]} &= N\left(\beta^{[M]} + 1 - \frac{1}{Q_\infty^{[M]}}\left(\frac{1}{\ln M} + {\alpha^{[M]}}^2 - \alpha^{[M]}\right)\right) \\
&\quad + N\delta^{[M]}(N) \\
&\quad + O\left(N^{\frac{1}{2}}\right)
\end{aligned}
$$

content

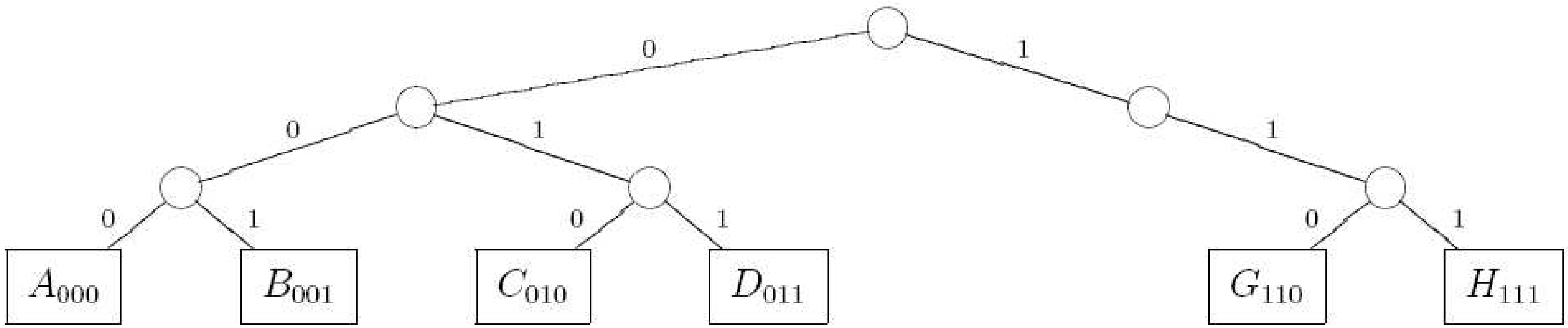
- ➪ **Tree**
- ➪ **Digital search tree**
- ➠ **Tries**
  - Defintions
  - Average case analysis
- ➪ **General framework**

content

- **Tree**
- **Digital search tree**
- **Tries:**
- **Defintion of**
  - – Digital search trie
  - – Patricia trie
- Average case analysis
- **General framework**

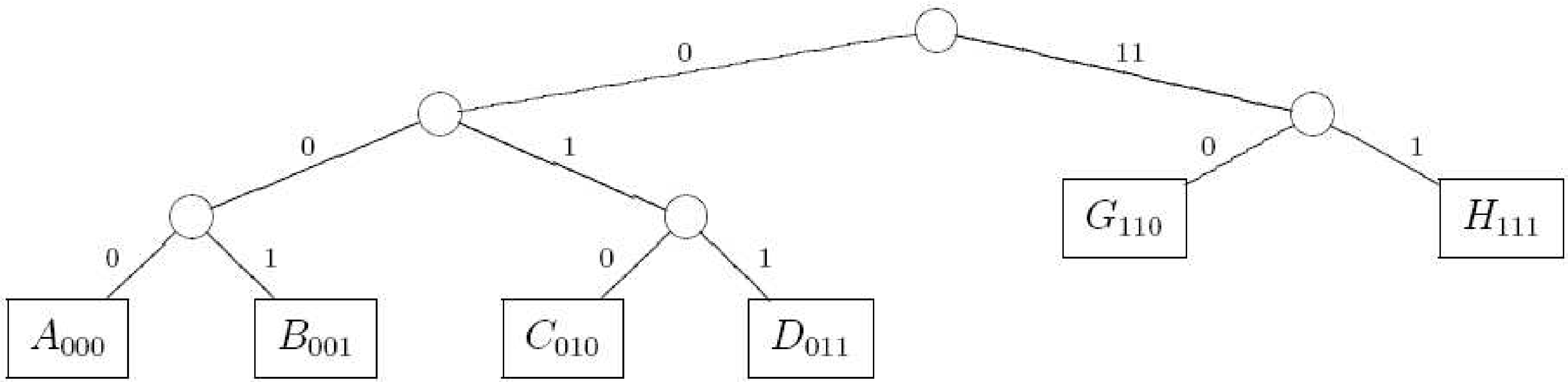
Digital search trie

A **digital search trie** is a digital tree for storing a set of strings in which there is one node for every prefix of every string in the set.



Patricia trie

A **Patricia tree** is defined as a compact representation of a digital search trie where all nodes with one child are merged with their parent.



content

- **Tree**
- **Digital search tree:**
- **Tries:**
  - Defintions
- **Average case analysis:**
  - – External path length
  - – External internal nodes
- **General framework**

content

- ➪ **Tree**
- ➪ **Digital search tree**
- ➪ **Tries:**
  - Defintions
  - Average case analysis:

    ➠ **External path length**

    – External internal nodes
- ➪ **General framework**

External path length for digital search trie

Fundamental recurrence relation

$$A_N^{[T]} = N + \sum_{k=0}^{\infty} \frac{1}{2^N} \binom{N}{k} \left(A_k^{[T]} + A_{N-k}^{[T]} \right), \qquad N \geq 2$$

with $A_0^{[T]} = A_1^{[T]} = 0$

External path length for digital search trie

Transformation

$$A^{[T]}(z) = z\left(e^{z} - 1\right) + 2A^{[T]}\left(\frac{z}{2}\right)e^{z-2}$$

External path length for digital search trie

Substitution by $B(z)$

$$A(z) = e^{z} B(z)$$

$$B^{[T]}(z) = z\left(1 - e^{-z}\right) + 2B^{[T]}\left(\frac{z}{2}\right)$$

External path length for digital search trie

Explicit formula

$$B^{[T]}(z) = \frac{N(-1)^N}{1 - \left(\frac{1}{2}\right)^{N-1}}$$

$$A_N^{[T]} = \sum_{k=2}^{\infty} \binom{N}{k} \frac{k(-1)^k}{1 - \left(\frac{1}{2}\right)^{k-1}}$$

External path length for digital search trie

average case

$$A_N^{[T]} = N \lg N + N\left(\frac{\gamma}{\ln 2} + \frac{1}{2} + \delta\left(N\right)\right) + O(1)$$

External path length for Patricia trie

Fundamental recurrence relation

$$A_N^{[P]} = N\left(1 - \frac{1}{2^{N-1}}\right) + \sum_{k=0}^{\infty} \frac{1}{2^N}\binom{N}{k}\left(A_k^{[P]} + A_{N-k}^{[P]}\right), \qquad N \geq 1$$

External path length for Patricia trie

Transformation

$$A^{[P]}(z) = z\left(e^{z} - e^{\frac{z}{2}}\right) + 2A^{[P]}\left(\frac{z}{2}\right)e^{\frac{z}{2}}$$

External path length for Patricia trie

Substitution

$$B^{[P]}(z) = z\left(1 - e^{-\frac{z}{2}}\right) + 2B^{[P]}\left(\frac{z}{2}\right)$$

$$B^{[P]}(z) = \frac{N(-1)^N}{2^{N-1} - 1}$$

External path length for Patricia trie

Explicit formula

$$A_N^{[P]} = \sum_{k=2}^{\infty} \binom{N}{k} \frac{k\,(-1)^k}{2^{k-1}-1} = A_N^{[T]} - N$$

content

- **Tree**
- **Digital search tree**
- **Tries:**
  - Defintions
  - Average case analysis:
    - External path length
- **External internal nodes**
- **General framework**

External internal nodes for Patricia trie

Fundamental recurrence relation

$$C_N^{[P]} = \sum \frac{1}{2^N} \binom{N}{k} \left(C_k^{[P]} + C_{N-k}^{[P]} \right), \qquad N \geq 3$$

with $C_0^{[P]} = C_1^{[P]} = 0$ and $C_2^{[P]} = 1$

External internal nodes for Patricia trie

Transformation

$$C^{[P]}(z) = \left(\frac{z}{2}\right)^2 + 2C^{[P]}\left(\frac{z}{2}\right)e^{\frac{z}{2}}$$

External internal nodes for Patricia trie

Substitution

$$D^{[P]}(z) = \left(\frac{z}{2}\right)^2 e^{-z} + 2D^{[P]}\left(\frac{z}{2}\right)$$

External internal nodes for Patricia trie

Explicit formula

$$C_N^{[P]} = \frac{1}{4}\sum_{k=2}^{N}\binom{N}{k}\frac{k\,(k-1)\,(-1)^k}{1-\frac{1}{2}^{k-1}}$$

External internal nodes for Patricia trie

average case

$$C_N^{[P]} = N\left(\frac{1}{4\ln 2} + \overline{\delta}^{[P]}(N)\right)$$

content

- Tree
- Digital search tree
- Trie
- General framework

General Framework for digital search trees

Fundamental recurrence relation

$$X(T) = \sum_{subtrees\,T_j\,of\,the\,root\,of\,T} X(T_j) + x(T)$$

General Framework for digital search trees

Transformation

$$X(z) = \sum_{N=0}^{\infty} X_N \frac{z^N}{N!}$$

$$X'(z) = MX\left(\frac{z}{M}\right) e^{\left(1-\frac{1}{M}\right)z} + x(z)$$

General Framework for digital search trees

Substitution

$$Y(z) = e^{-z} X(z)$$

$$y(z) = e^{-z} x(z)$$

$$Y'(z) + Y(z) = MY\left(\frac{z}{M}\right) + y'(z) + y(z)$$

General Framework for digital search trees

Explicit formula

$$X_N = \sum_{k=0}^{N} \binom{N}{k} Y_k$$

General Framework for digital search trees

Asymptotic analysis of $(-1)^k Y_k$

Find a function Y_k^* which

(i) is simply related to Y_k so that $\sum_{k=0}^{N} \binom{N}{k} \left(Y_k - (-1)^k Y_k^*\right)$ is easily evaluated,

(ii) satisfies a recurrence of the form
$Y_{N+1}^* = (1 - g(M, N))\, Y_N^* + f(M, N)$,

(iii) goes to zero quickly as $N \to \infty$.

General Framework for digital search trees

Asymptotic analysis of $(-1)^k Y_k$

Turn the recurrence around to extend Y_N^* to the complex plane.

Evaluate $\sum_{k=0}^{N} \binom{N}{k} \left(Y_k - (-1)^k Y_k^*\right)$ as detailed in the previous sections.

General Framework for tries

$$X(T) = \sum_{subtrees\, T_j\, of\, the\, root\, of\, T} X(T_j) + x(T)$$

$$X(z) = MX\left(\frac{z}{M}\right)e^{\frac{z}{M}} + x(z)$$

This can be solved by Rice's method or also by Mellin transform techniques.

Thank you for your attention!