Probabilistic analysis of a search tree problem

Henning Sulzbach
J. W. Goethe-Universität Frankfurt a. M.

XII Latin American Congress of Probability and Mathematical Statistics

Viña del Mar, March 27, 2012

joint work with Nicolas Broutin and Ralph Neininger

Partial match retrieval

A classical combinatorial problem is to perform a search in a multidimensional database where the record to be retrieved is either *fully* or *partially* specified. The latter is called a *Partial match query*.

n-dim. domain: $S = S_1 \times \cdots \times S_n$

set of data $S' \subseteq S$ with $|S'| < \infty$.

Problem: For a fixed query $q = (q_1, \ldots, q_n)$ with $q_i \in S_i \cup \{*\}$, find all elements $s = (s_1, \ldots, s_n) \in S'$ such that

$$s_i = q_i, \qquad \text{if} \qquad q_i \neq *.$$

Data structures

Comparison-based structures - search trees:

- Quadtrees (Finkel and Bentley '74),
- K-d-trees (Bentley '75)

Several variants are known in the literature.

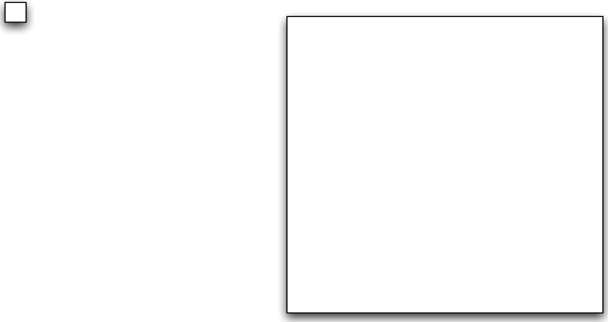
Digital structures:

- K-d-tries (Rivest '76)

The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Dimension: $n = 2$

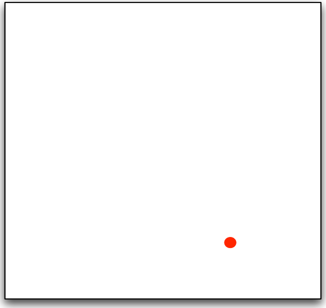


The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

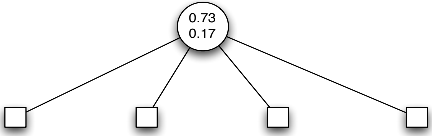
$0.73, 0.17$

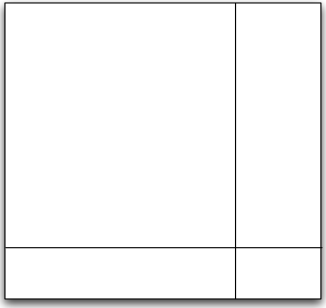


The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$



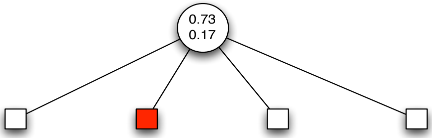


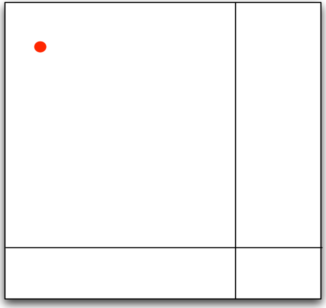
The Quadtree - Construction

Model: $S_i = [0,1]$ for all i.

Quadtree: $n = 2$

$0.1, 0.86$

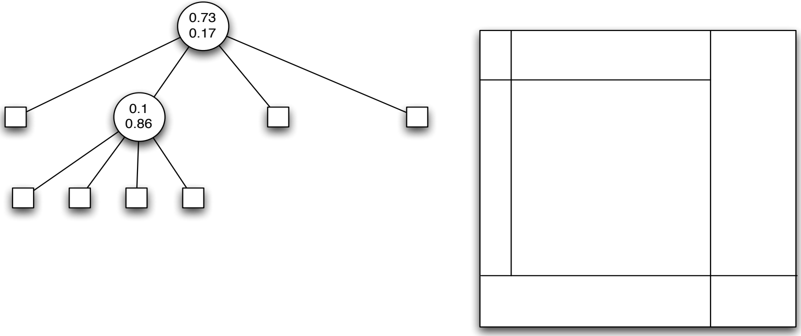




The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

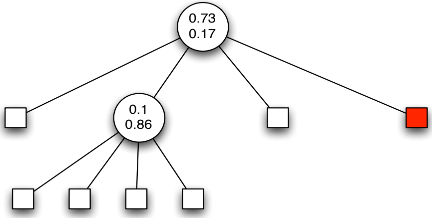


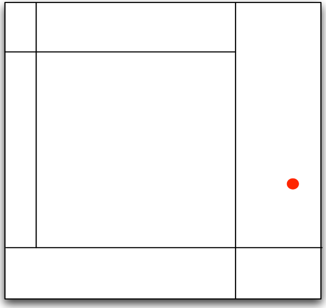
The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

$0.93, 0.36$

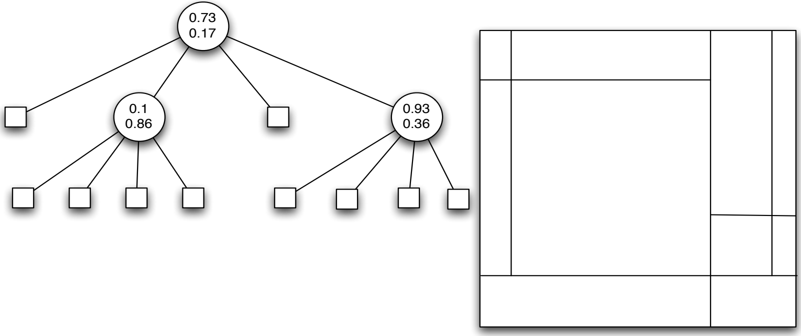




The Quadtree - Construction

Model: $S_i = [0,1]$ for all i.

Quadtree: $n = 2$

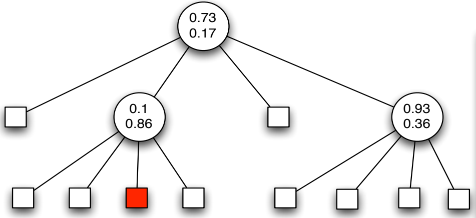


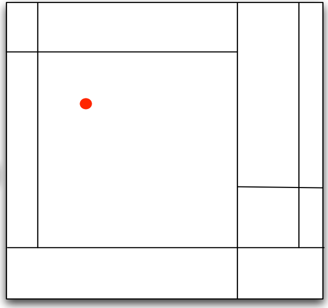
The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

$0.26, 0.64$

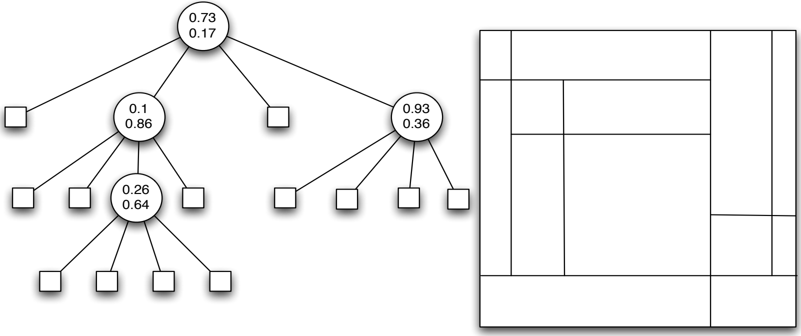




The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

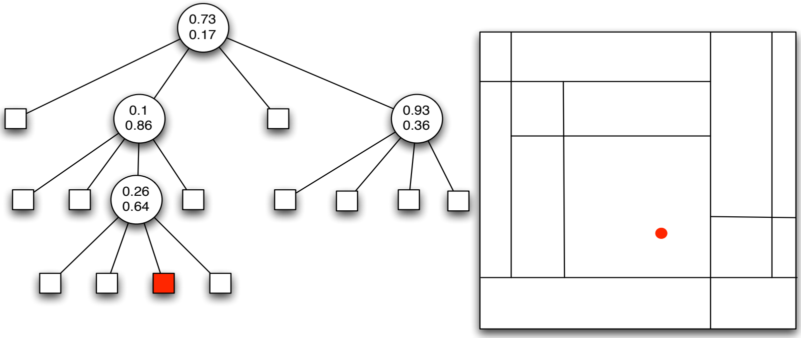


The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

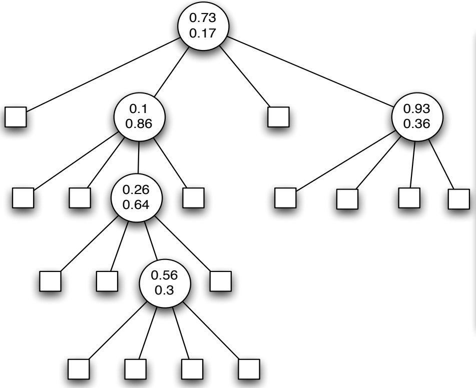
$0.56, 0.3$

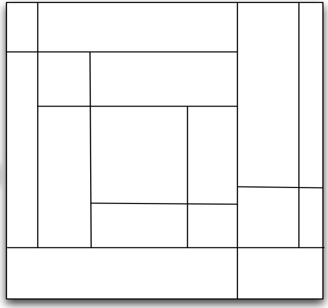


The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$



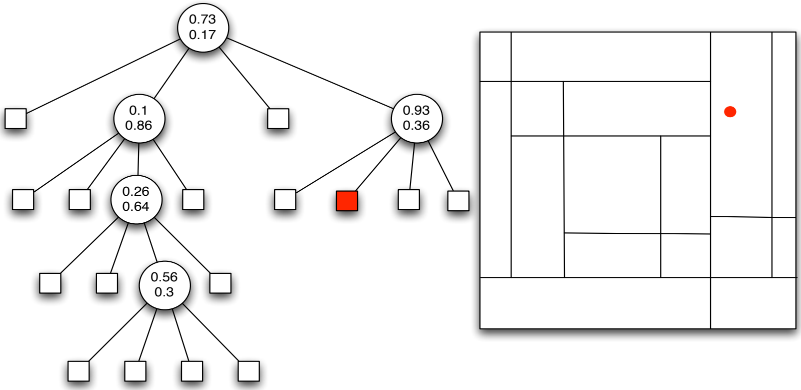


The Quadtree - Construction

Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$

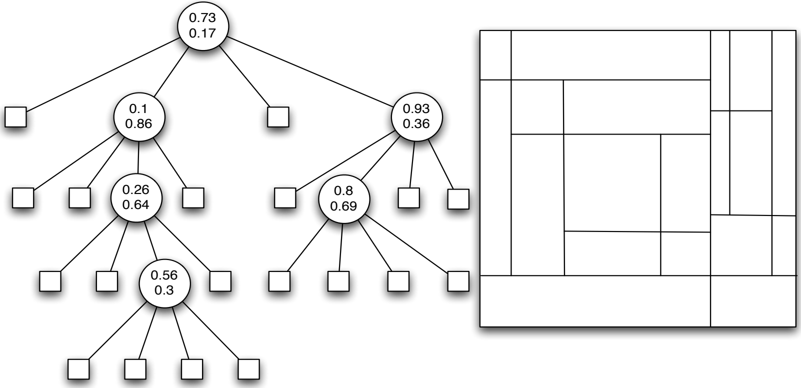
$0.8, 0.69$



The Quadtree - Construction

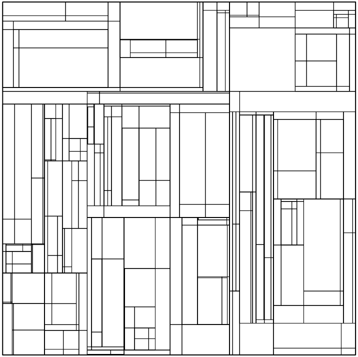
Model: $S_i = [0, 1]$ for all i.

Quadtree: $n = 2$



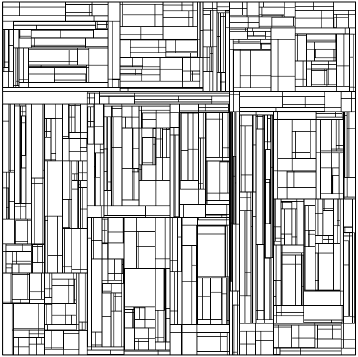
Simulation - Quadtree

$n = 100$



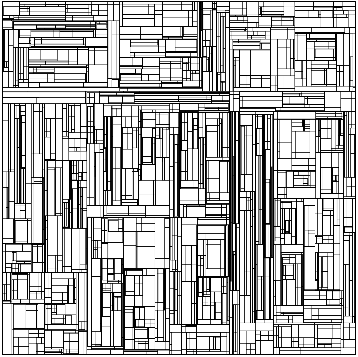
Simulation - Quadtree

$n = 500$



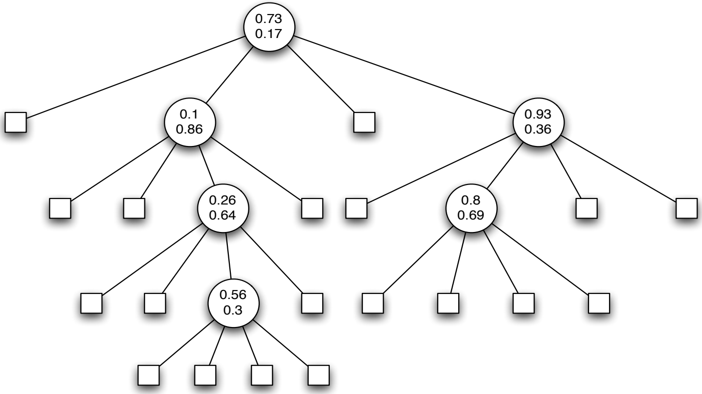
Simulation - Quadtree

$n = 1000$



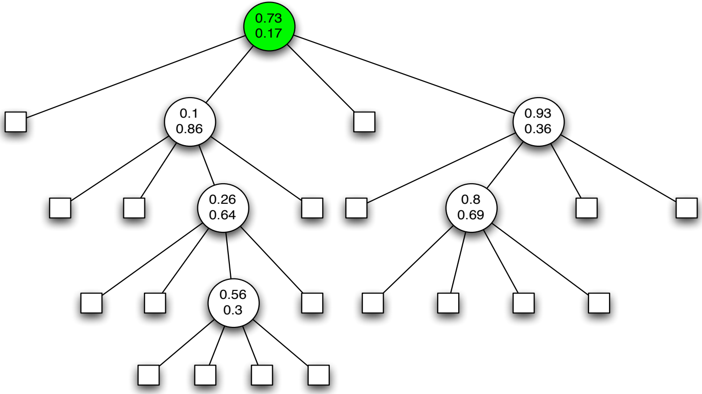
A partial match query

Query: $q = \{s, *\}$, $\quad s = 0.2$



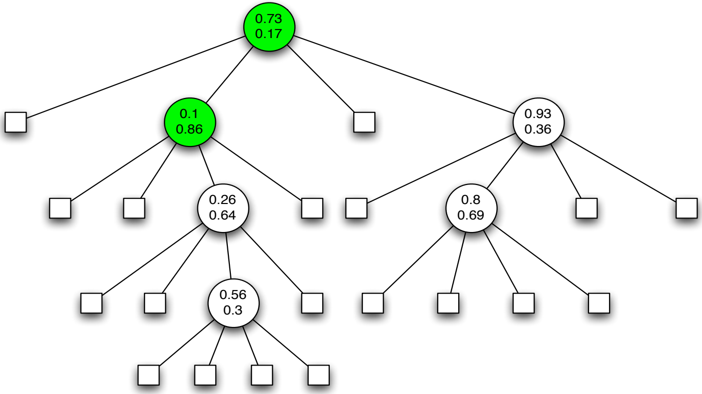
A partial match query

Query: $q = \{s, *\}$, $\quad s = 0.2$



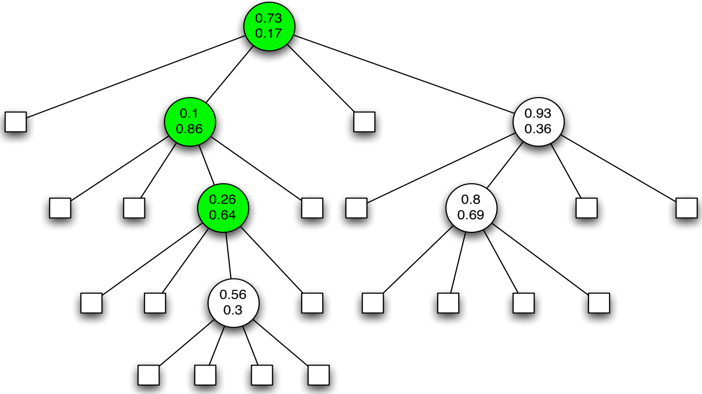
A partial match query

Query: $q = \{s, *\}$, $\quad s = 0.2$



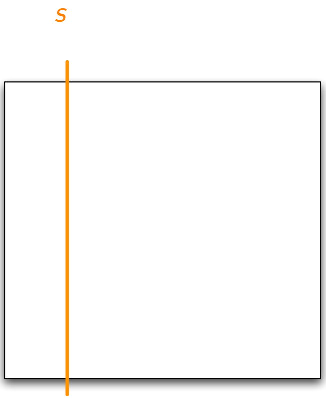
A partial match query

Query: $q = \{s, *\}$, $\quad s = 0.2$



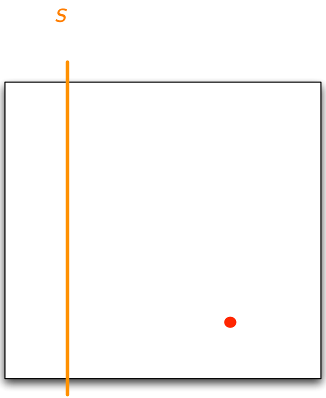
A partial match query

Query: $q = \{s, *\}$, $s = 0.2$



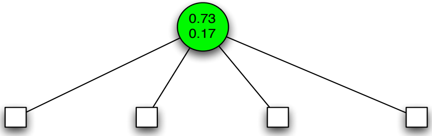
A partial match query

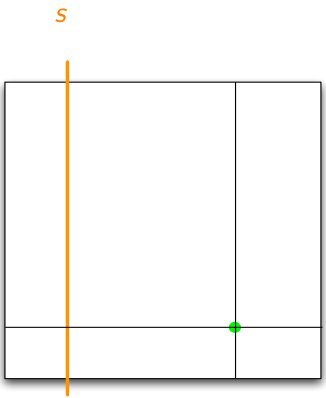
Query: $q = \{s, *\}$, $s = 0.2$



A partial match query

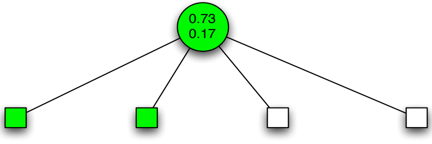
Query: $q = \{s, *\}$, $s = 0.2$

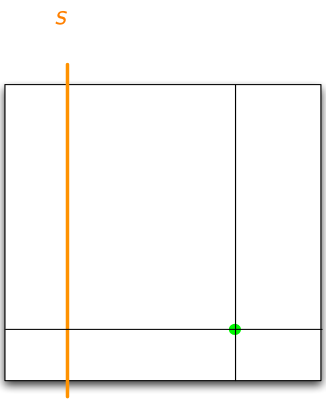




A partial match query

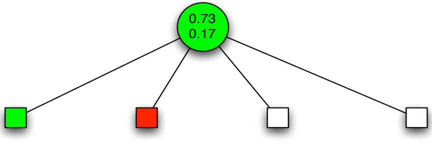
Query: $q = \{s, *\}$, $s = 0.2$

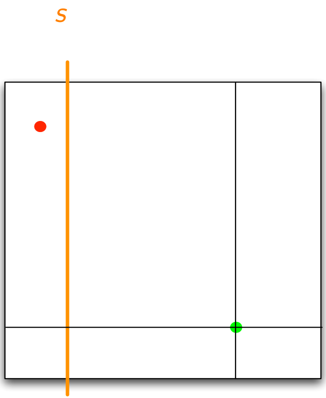




A partial match query

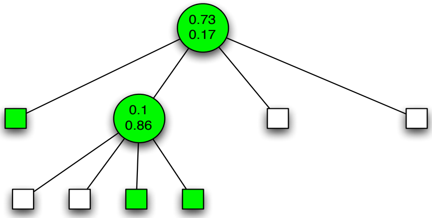
Query: $q = \{s, *\}$, $s = 0.2$

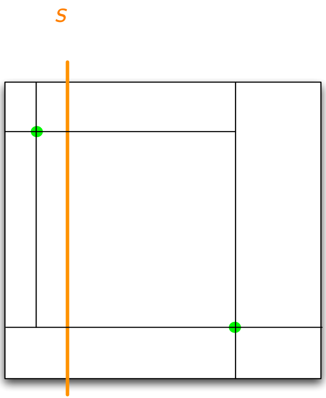




A partial match query

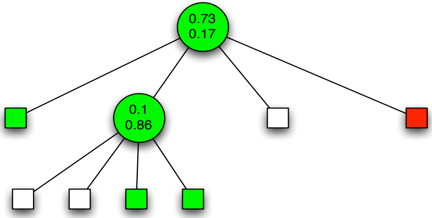
Query: $q = \{s, *\}$, $s = 0.2$

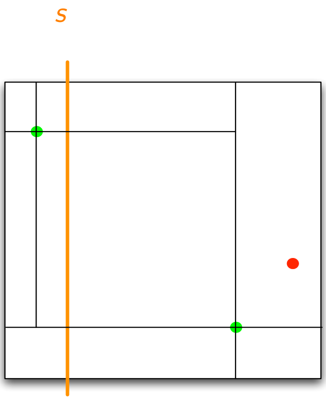




A partial match query

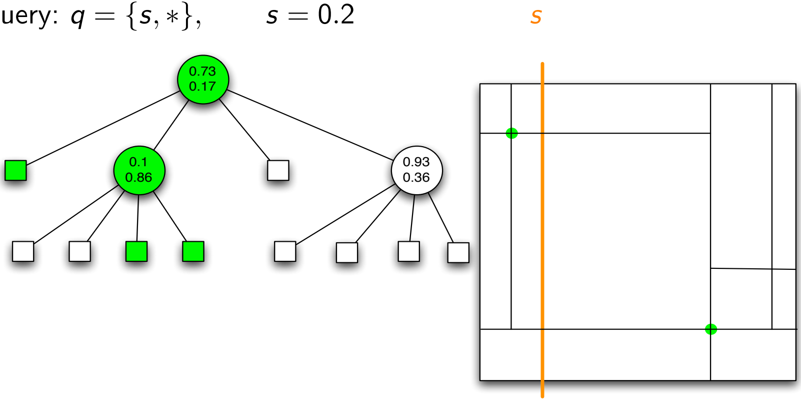
Query: $q = \{s, *\}$, $s = 0.2$





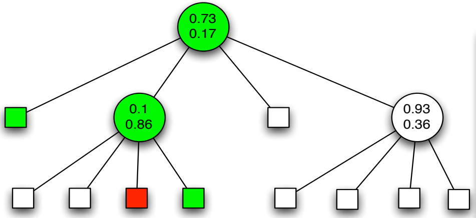
A partial match query

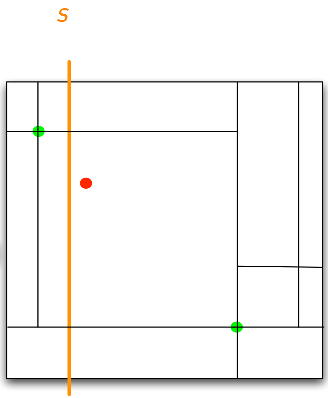
Query: $q = \{s, *\}$, $s = 0.2$



A partial match query

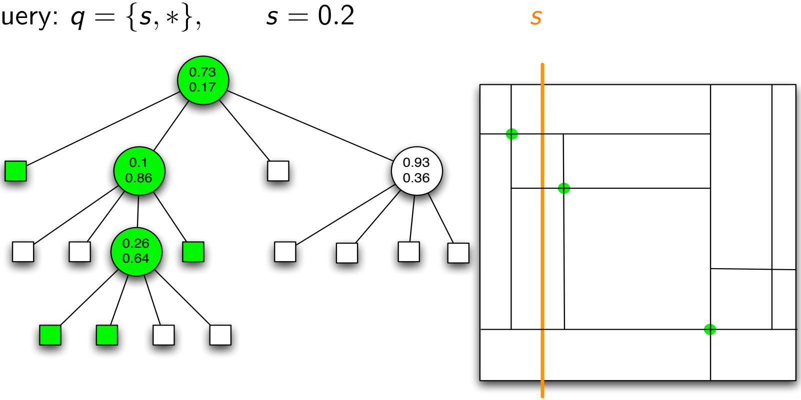
Query: $q = \{s, *\}$, $s = 0.2$





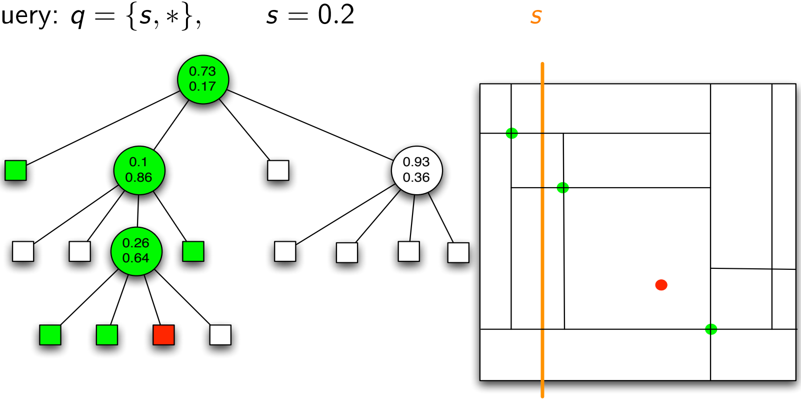
A partial match query

Query: $q = \{s, *\}$, $s = 0.2$



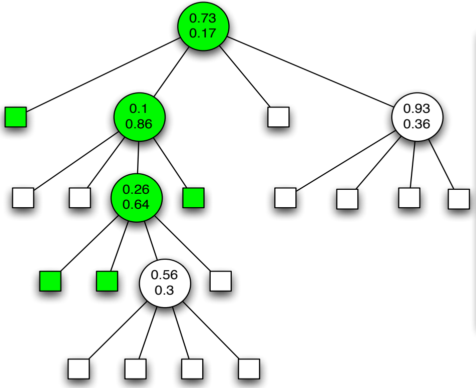
A partial match query

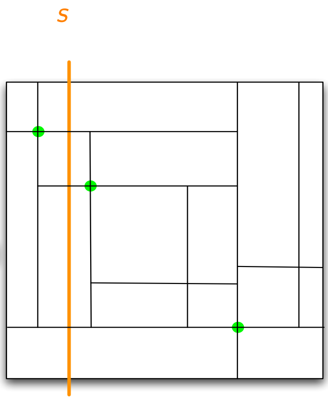
Query: $q = \{s, *\}$, $s = 0.2$



A partial match query

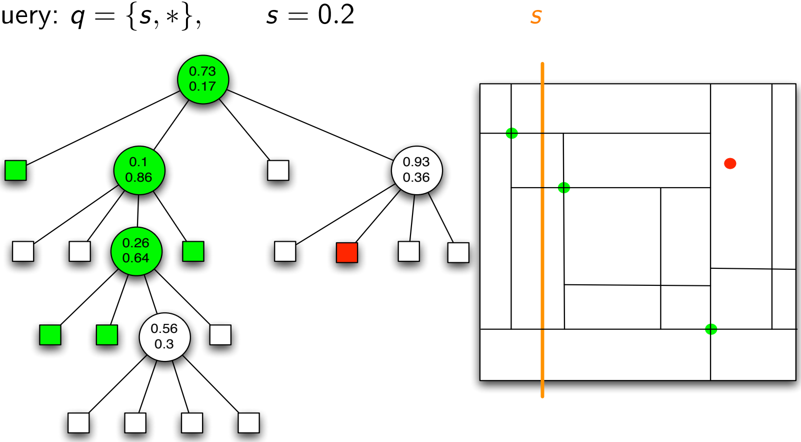
Query: $q = \{s, *\}$, $s = 0.2$





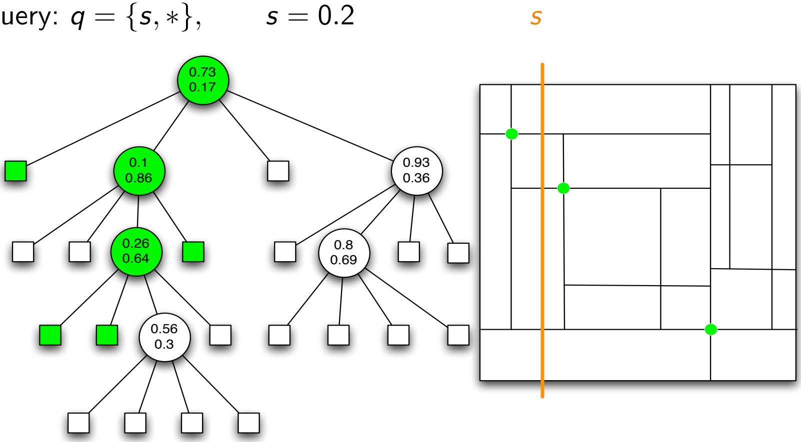
A partial match query

Query: $q = \{s, *\}$, $s = 0.2$



A partial match query

Query: $q = \{s, *\}$, $s = 0.2$



A problem of stochastic geometry

Performing a partial match query with $q = \{s, *\}$, a node is visited *if and only if* it is inserted in a subregion that intersects the vertical line $x = s$.

This is equivalent to an intersection of its horizontal line and the line $x = s$.

Probabilistic model

For the analysis of the complexity of information retrieval we *always* assume the components of elements in the database S' to be *independent* and *uniform* on $[0, 1]^2$.

$C_n(s)$: number of nodes visited by a partial match query with $q = \{s, *\}$ in a random two-dimensional quadtree of size n.

Probabilistic analysis of the complexity

Theorem (Flajolet, Gonnet, Puech, Robson '93)

Let ξ *be uniform on* $[0,1]$, *independent of the quadtree. For* $n \to \infty$, *it holds*

$$\mathbb{E}[C_n(\xi)] \sim \kappa n^{\beta}$$

with

$$\kappa = \frac{\Gamma(2\beta+2)}{2\Gamma^3(\beta+1)} \approx 1.59, \qquad \beta = \frac{\sqrt{17}-3}{2} \approx 0.56.$$

The variance or a distributional limit theorem remained open problems.

Asymptotic results for fixed s

Theorem (Curien, Joseph '11)

For fixed $s \in [0,1]$ and $n \to \infty$, it holds

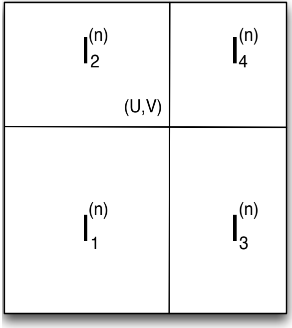
$$\mathbb{E}C_n(s) \sim K_1 n^\beta (s(1-s))^{\beta/2},$$

where

$$K_1 \int_0^1 (s(1-s))^{\beta/2} ds = \kappa.$$

The main idea - Decomposing at the root

U, V : components of the first inserted point,
$I_1^{(n)}, \ldots, I_4^{(n)}$: number of points in the subregions.



Given U, V, we have

$$\mathcal{L}(I_1^{(n)}, I_2^{(n)}, I_3^{(n)}, I_4^{(n)}) = \text{Mult}(n-1; UV, U(1-V), (1-U)V, (1-U)(1-V)).$$

The main idea - Decomposing at the root

For any $s \in [0,1]$,

$$\begin{aligned} C_n(s) \quad &\stackrel{d}{=} \quad 1 + 1_{\{s<U\}} \left(C^{(1)}_{I_1^{(n)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_2^{(n)}}\left(\frac{s}{U}\right) \right) \\ &\quad + 1_{\{s \geq U\}} \left(C^{(3)}_{I_3^{(n)}}\left(\frac{s-U}{1-U}\right) + C^{(4)}_{I_4^{(n)}}\left(\frac{s-U}{1-U}\right) \right), \end{aligned}$$

where $(C_n^{(1)}), (C_n^{(2)}), (C_n^{(3)}), (C_n^{(4)})$ are ind. copies of (C_n), ind. of $(U, V, I_1^{(n)}, I_2^{(n)}, I_3^{(n)}, I_4^{(n)})$.

This does not imply a recurrence for $C_n(s)$, neither for fixed s nor for $s = \xi$. It is due to this fact that the problem remained unsolved for many years.

The recursion on the process level

The recursion

$$C_n(s) \stackrel{d}{=} 1 + 1_{\{s<U\}} \left(C^{(1)}_{I_1^{(n)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_2^{(n)}}\left(\frac{s}{U}\right) \right) + 1_{\{s \geq U\}} \left(C^{(3)}_{I_3^{(n)}}\left(\frac{s-U}{1-U}\right) + C^{(4)}_{I_4^{(n)}}\left(\frac{s-U}{1-U}\right) \right),$$

remains valid on the level of càdlàg functions, $(C_n(s))_{s\in[0,1]}$ is a random stepfunction!

The recursion on the process level

Scaling gives

$$\frac{C_n(s)}{n^\beta} \stackrel{d}{=} n^{-\beta} + 1_{\{s<U\}} \left(\left(\frac{I_1^{(n)}}{n}\right)^\beta \frac{C^{(1)}_{I_1^{(n)}}\left(\frac{s}{U}\right)}{\left(I_1^{(n)}\right)^\beta} + \left(\frac{I_2^{(n)}}{n}\right)^\beta \frac{C^{(2)}_{I_2^{(n)}}\left(\frac{s}{U}\right)}{\left(I_2^{(n)}\right)^\beta} \right)$$

$$+1_{\{s\geq U\}} \left(\left(\frac{I_3^{(n)}}{n}\right)^\beta \frac{C^{(3)}_{I_3^{(n)}}\left(\frac{s-U}{1-U}\right)}{\left(I_3^{(n)}\right)^\beta} + \left(\frac{I_4^{(n)}}{n}\right)^\beta \frac{C^{(4)}_{I_4^{(n)}}\left(\frac{s-U}{1-U}\right)}{\left(I_4^{(n)}\right)^\beta} \right).$$

Fixed-point equation

Assuming $n^{-\beta} C_n(s) \to Z(s)$ uniformly in $s \in [0,1]$ for $n \to \infty$, suggests that Z satisfies

$$
\begin{aligned}
Z(s) \quad &\stackrel{d}{=} \quad 1_{\{s<U\}} \left((UV)^\beta Z^{(1)} \left(\frac{s}{U} \right) + (U(1-V))^\beta Z^{(2)} \left(\frac{s}{U} \right) \right) \\
&\quad +1_{\{s \geq U\}} ((1-U)V)^\beta Z^{(3)} \left(\frac{s-U}{1-U} \right) \\
&\quad +1_{\{s \geq U\}} ((1-U)(1-V))^\beta Z^{(4)} \left(\frac{s-U}{1-U} \right),
\end{aligned}
$$

where $Z^{(1)}, Z^{(2)}, Z^{(3)}, Z^{(4)}$ are ind. copies of Z, ind. of (U, V).

A functional limit law

Theorem (Broutin, Neininger, S. '12)

There exists a random continuous process Z on the unit interval such that

$$\left(\frac{C_n(s)}{K_1 n^{\beta}}\right)_{s\in[0,1]} \to (Z(s))_{s\in[0,1]}, \quad n \to \infty,$$

in distribution in $(\mathcal{D}[0,1], d_{sk})$ where d_{sk} denotes the Skorohod metric.

Characterization of Z

Theorem (Broutin, Neininger, S. '12)

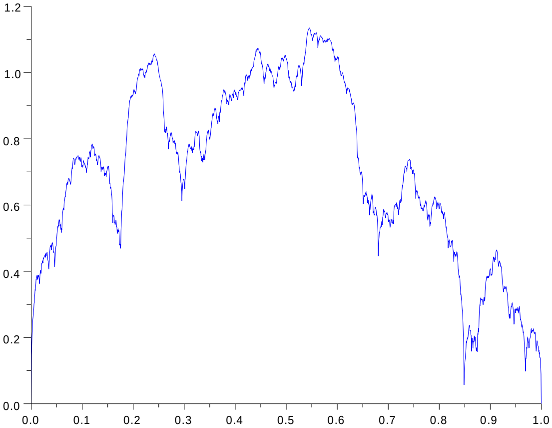
Z is the unique solution in $(\mathcal{D}[0,1], d_{sk})$ of the fixed-point equation

$$\begin{aligned} Z(s) \quad &\stackrel{d}{=} \quad 1_{\{s<U\}} \left((UV)^\beta Z^{(1)} \left(\frac{s}{U} \right) + (U(1-V))^\beta Z^{(2)} \left(\frac{s}{U} \right) \right) \\ &\quad +1_{\{s \geq U\}} ((1-U)V)^\beta Z^{(3)} \left(\frac{s-U}{1-U} \right) \\ &\quad +1_{\{s \geq U\}} ((1-U)(1-V))^\beta Z^{(4)} \left(\frac{s-U}{1-U} \right), \end{aligned}$$

with $\mathbb{E}\|Z\|^2 < \infty$ *and* $\mathbb{E}Z(\xi) = B(\beta/2+1, \beta/2+1)$.
Here, $Z^{(1)}, Z^{(2)}, Z^{(3)}, Z^{(4)}$ *are independent copies of* Z*, independent of* (U, V)*.*

A Simulation

by Nicolas Broutin



The marginals of Z

Theorem (Broutin, Neininger, S. '12)

For all $s \in [0,1]$, we have

$$Z(s) \stackrel{d}{=} Z \cdot (s(1-s))^{\beta/2},$$

where Z is the unique solution of

$$Z \stackrel{d}{=} V^{\beta} U^{\beta/2} Z + (1-V)^{\beta} U^{\beta/2} Z'$$

with $\mathbb{E} Z = 1$ and $\mathbb{E} Z^2 < \infty$. Again, Z' is an independent copy of Z and (Z, Z') is independent of (U, V).

Back to the uniform case

Theorem (Broutin, Neininger, S. '12)

We have

$$\frac{C_n(\xi)}{\kappa n^\beta} \xrightarrow{d} Z \cdot \frac{(\xi(1-\xi))^{\beta/2}}{B(\beta/2+1, \beta/2+1)}$$

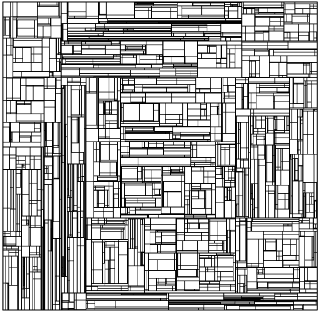
with convergence of all moments , in particular

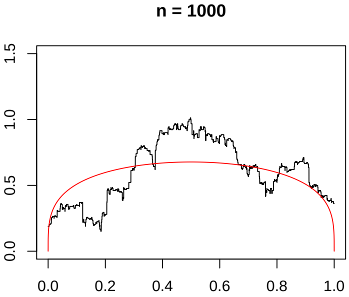
$$Var[C_n(\xi)] \sim K_2 n^{2\beta},$$

where

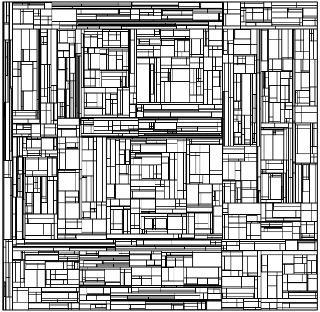
$$K_2 = K_1^2 \left[\frac{2(2\beta+1)}{3(1-\beta)} B^2(\beta+1, \beta+1) - B^2\left(\frac{\beta}{2}+1, \frac{\beta}{2}+1\right)\right] = 0.44736\ldots$$

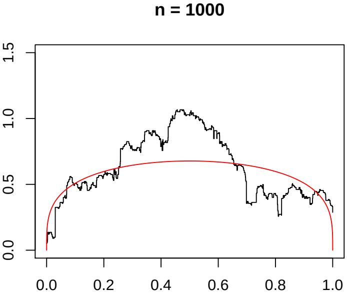
Simulations



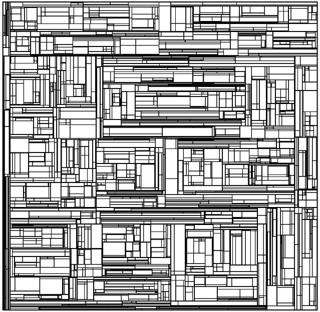


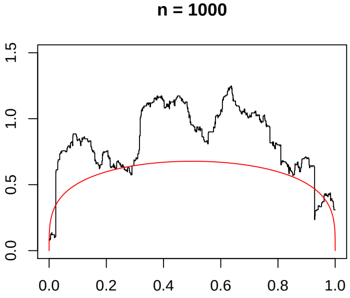
Simulations





Simulations





The proof - Functional contraction method

Solutions to the fixed-point equation of interest, or more generally of type

$$Z \stackrel{d}{=} \sum_{i=1}^{K} A_r Z_r + b,$$

with conditions as in our case and random linear operators $A_1, \ldots, A_K$ are considered as fixed-points of the map

$$\begin{aligned} T : \mathcal{M}(\mathcal{D}[0,1]) &\to \mathcal{M}(\mathcal{D}[0,1]), \\ T(\mu) &= \mathcal{L}\left(\sum_{i=1}^{K} A_r Z_r + b\right), \end{aligned}$$

where $Z_1, \ldots, Z_r$ are independent with common distribution μ, independent of $(A_1, \ldots, A_K, b)$.

The proof - Functional contraction method

- Choose a suitable subset of $\mathcal{M}(\mathcal{D}[0,1])$ and endow it with some appropriate metric d that turns T into a contraction. Here, the crucial condition turns out to be

$$\sum_{l=1}^{K} \mathbb{E}\|A_i\|_{\mathrm{op}}^s < 1,$$

  for $s < 1$.
- Construct a solution of the fixed-point equation by hand.
- Show $d(C_n^*, Z) \to 0$ and infer distributional convergence for the rescaled quantity C_n^*.

The why of β - Size-biasing!

Let $X_n = C_n(\xi)$. On the level of expectations,

$$\mathbb{E}[X_n] \quad = \quad 1 + 2\mathbb{E}\left[1_{\{\xi<U\}} X^{(1)}_{I_1^{(n)}} + 1_{\{\xi\geq U\}} X^{(3)}_{I_3^{(n)}}\right].$$

This allows to compute β:

$$\mathbb{E}[X_n] \quad = \quad 1 + 2\mathbb{E}[X_{L_n}]$$

with $L_n \stackrel{d}{=} \mathrm{Bin}(n-1, \sqrt{U}V)$. Scaling gives

$$n^{-\gamma}\mathbb{E}[X_n] \sim 2\mathbb{E}\left[\left(\frac{L_n}{n}\right)^\gamma \frac{X_{L_n}}{L_n^\gamma}\right].$$

Hence $1 = 2\mathbb{E}[(\sqrt{U}V)^\gamma] \Rightarrow \gamma = \beta$.

The constant β appears in several other contexts, e.g. as the Hausdorff dimension of the random Cantor set.

References

- Flajolet, Gonnet, Puech, Robson. Analytic variations on quadtrees. *Algorithmica*, 10:472–500, 1993.
- Curien, Joseph. Partial match queries in 2-dimensional quadtrees: a probabilistic approach. *Adv. Appl. Probab.*, 43(1):178–194, 2011.
- Broutin, Neininger, S. Partial match queries in random quadtrees. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1056–1065, 2012.
- Neininger, S. On a functional contraction method, *submitted*, 2012, available at http://arxiv.org.
- Broutin, Neininger, S. A limit process for partial match queries in random quadtrees, *submitted*, 2012, available at http://arxiv.org.