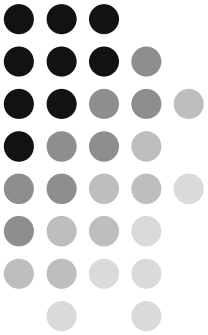
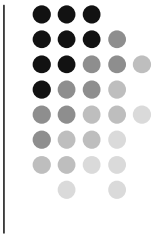
Vyhledávání v multimediálních datech

Vlastislav Dohnal





Osnova

- úvod
- podobnostní hledání
- podobnostní dotazy
- základní přístupy
- M-strom, D-Index

Úvod

- podobnostní hledání založené na relativních vzdálenostech
- příklady dat – množiny, vektory, dokumenty
- problém – najít objekty podobné dotazu
- metrické prostory
- velké archívy vyžadují podpůrné indexy pro urychlení vyhledávání

Metrický prostor

- metrický prostor $\mathcal{M} = (\mathcal{D}, d)$

  $\mathcal{D}$ – doména objektů
- vzdálenostní funkce $d(x,y)$

  $\forall x,y,z \in \mathcal{D}$
  - $d(x,y) \geq 0$
  - $d(x,y) = 0 \Leftrightarrow x = y$
  - $d(x,y) = d(y,x)$
  - $d(x,y) \leq d(x,z) + d(z,y)$

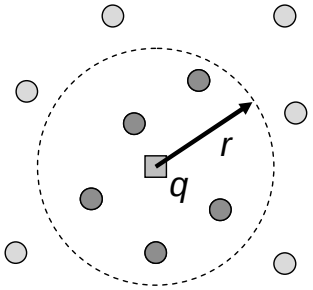
  - euklidovská (L_p metriky), editační, kvadratická, Hausdorffova vzdál., Jacardův koeficient, …

Vzdálenostní funkce

- L_p metrické funkce (pro vektory)
  - L_1 – městská vzdál. $L_1(x,y) = \sum_{i=1}^{n} | x_i - y_i |$
  - L_2 – euklidovská vzdál. $L_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
  - L_∞ – maximální vzdál. $L_\infty(x,y) = \max_{i=1}^{n} |x_i - y_i|$
- editační vzdálenost (pro řetězce)
  - minimální počet vložení, smazání a nahrazení jednoho znaku
  - d('application', 'applet') = 6
- Jacardův koeficient (pro množiny) $d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$

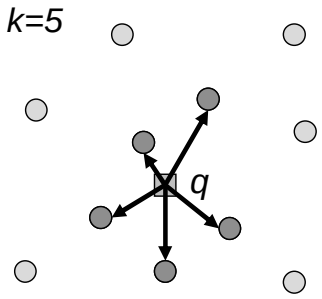
Podobnostní hledání

- Mějme $\mathcal{X} \subseteq \mathcal{D}$ v metrickém prostoru $\mathcal{M}$, předzpracuj $\mathcal{X}$ tak, že podobnostní dotazy budou řešeny efektivně.
- dvě fáze
  1. určení metriky
  2. vytvoření indexové struktury
- podobnostní dotazy
  - rozsahový dotaz
  - $R(q,r) = \{ x \in \mathcal{X} \mid d(q,x) \leq r \}$

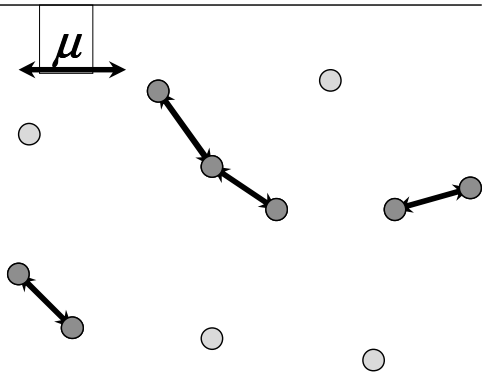


Podobnostní hledání

- k-nejbližších sousedů
  - $NN(q,k) = A$
  - $A \subseteq \mathcal{X}, |A| = k$
  - $\forall x \in A, y \in \mathcal{X} - A, d(q,x) \leq d(q,y)$

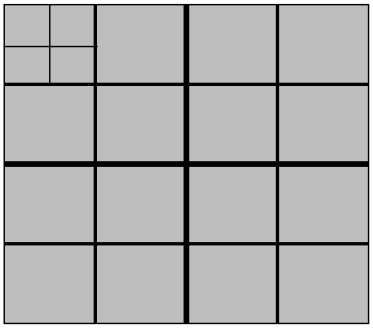


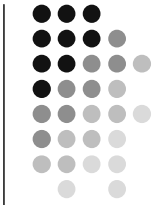
- podobnostní spojení
  - $X = \{x_1, x_2, \ldots x_N\}$, $Y = \{y_1, y_2, \ldots y_M\}$
  - $\{(x_i,y_j) \mid d(x_i,y_j) \leq \mu\}$
  - podobnostní „samo“ spojení $\Leftrightarrow$ X = Y



Základní přístupy

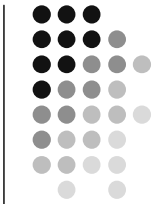
- vektorové prostory
  - mnoho indexových struktur
    - B^+ stromy, quad stromy, k-d stromy, R-stromy, X-stromy
  - problém transformace problému
    - jak reprezentovat řetězce
  - optimalizace V/V
  - vysoká dimenzionalita
    - neefektivnost struktur
    - lze použít redukční metody





Základní přístupy

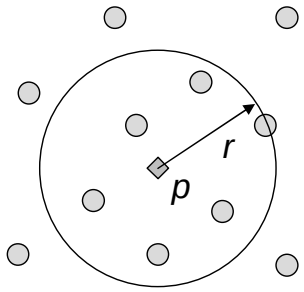
- metrické prostory
  - nevýhody:
    - žádná geometrická reprezentace
    - chybí souřadnicový systém
    - pouze relativní vzdálenosti mezi objekty
  - výhody:
    - obecné použití



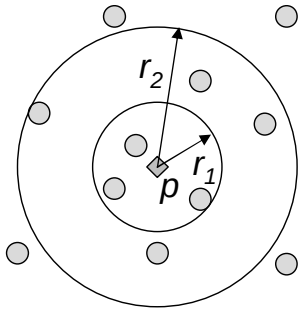
Indexování metrických prostorů

Základní principy

- dělení prostoru
  - sférické dělení
    - $\{ x \in \mathcal{X} \mid d(p,x) \leq r \}$
    - $\{ x \in \mathcal{X} \mid d(p,x) > r \}$

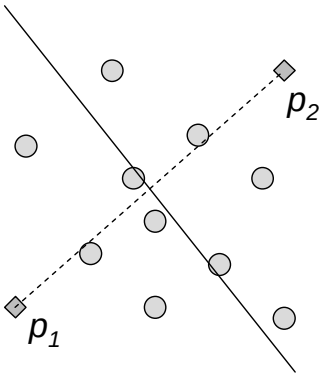


  - vícecestné sférické
    - $\{ x \in \mathcal{X} \mid d(p,x) \leq r_1 \}$
    - $\{ x \in \mathcal{X} \mid d(p,x) > r_1$ and $d(p,x) \leq r_2 \}$
    - $\{ x \in \mathcal{X} \mid d(p,x) > r_2 \}$



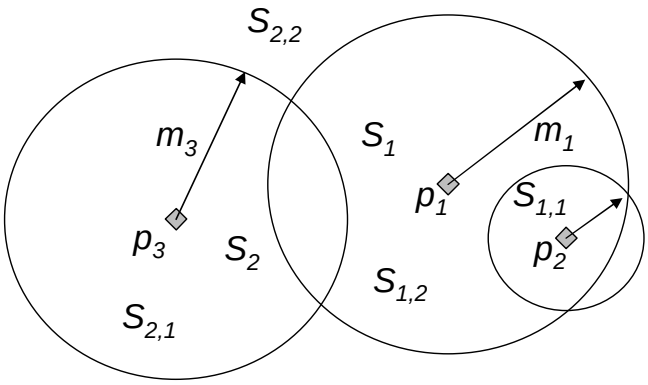
Základní principy

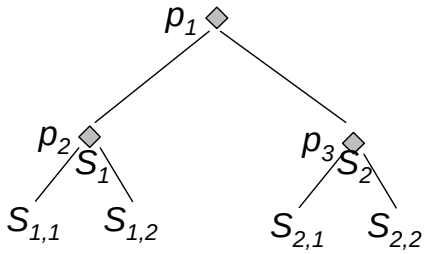
- dělení prostoru
  - hyper-rovinné dělení
    - $\{ x \in \mathcal{X} \mid d(p_1,x) \leq d(p_2,x) \}$
    - $\{ x \in \mathcal{X} \mid d(p_1,x) > d(p_2,x) \}$



Vantage Point Tree

- vyvážená binární stromová struktura
- sférické dělení
  - rekurzivně dělí datovou množinu X

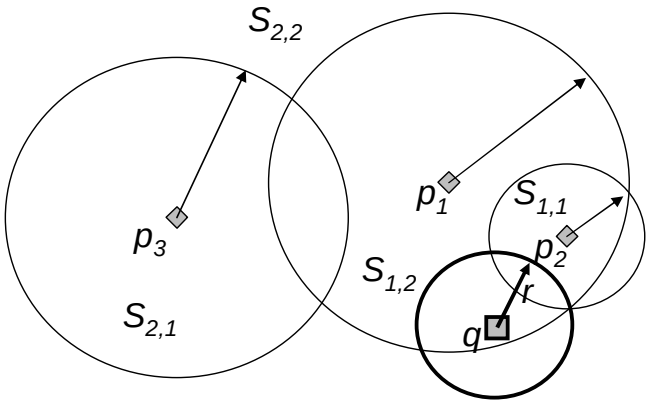


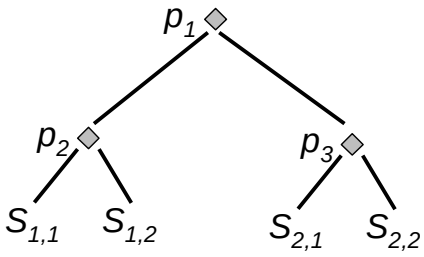


- statická struktura

Vantage Point Tree

Rozsahový dotaz *R(q,r)*:

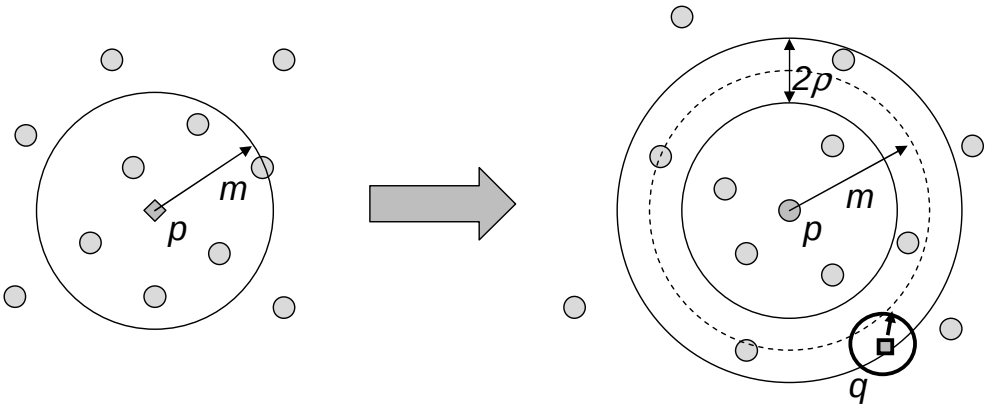




- if $d(q,p_i) - r \leq m_i$ prohledej levý podstrom
- if $d(q,p_i) + r \geq m_i$ prohledej pravý podstrom

Vantage Point Forest

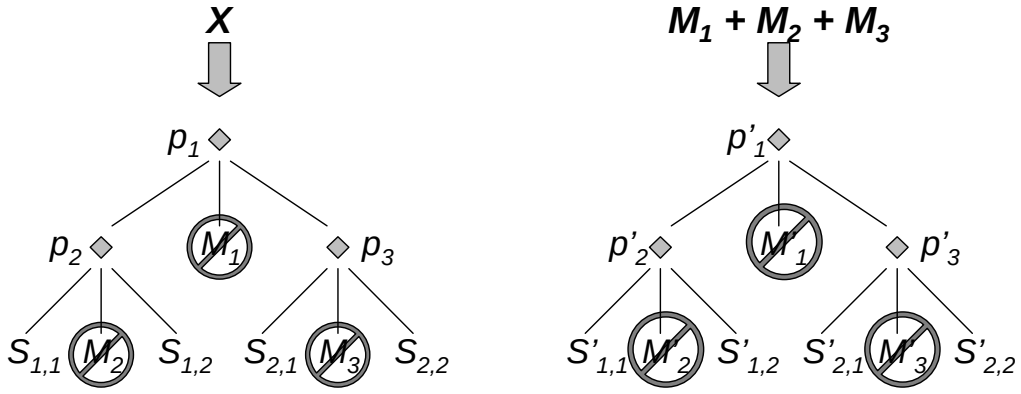
- les binárních stromů
- dělení s vyloučením
  - modifikované sférické dělení



- střední oblast je vyloučena z procesu dělení

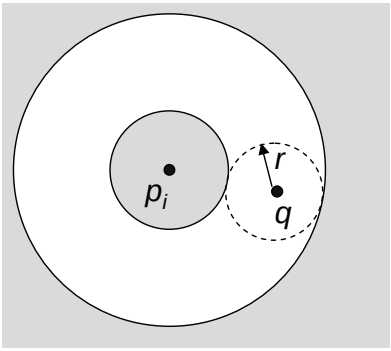
Vantage Point Forest

- rekurzivně je vybudován binární strom nad X
- oblasti vyloučení jsou použity pro další strom



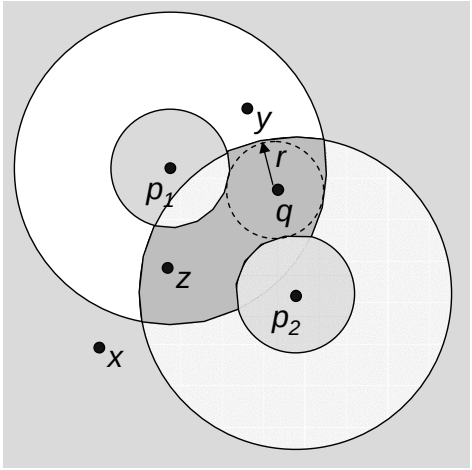
Základní principy

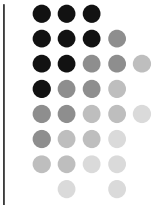
- filtrování
  - použití uložených vzdáleností k pivotům
  - založené na trojúhelníkové nerovnosti
  - vyřaď objekt $x \in \mathcal{X}$ pokud:
    - $d(p_i,x) < d(p_i,q) - r$
    - $d(p_i,x) > d(p_i,q) + r$



Filtrování

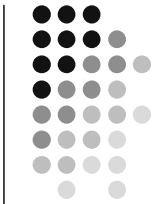
- mapování *F* do vektorového prostoru $(\mathbb{R}^t, L_\infty)$
- množina pivotů $T = \{ p_1, p_2, p_3, \ldots, p_t \}$
- $\forall x \in \mathcal{X}$, $F(x) = (d(x,p_1), \ldots, d(x,p_t))$
- vyřaď objekt *x*, pokud $L_\infty(F(x), F(q)) > r$
- *F* je kontraktivní
  - žádný vyloučený objekt nemůže být ve výsledku
  - některé vyhovující objekty nemusí být relevantní





Současné problémy

- statický návrh
  - vkládání a mazání objektů zhoršuje výkonnost při hledání
- operační paměť
  - zpracování velkých objemů dat není možné
- minimalizace CPU operací
  - minimalizace počtu volání vzdálenostní funkce
  - V/V operace nejsou uvažovány



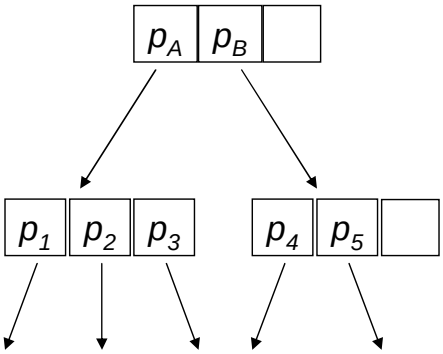
M-stromy

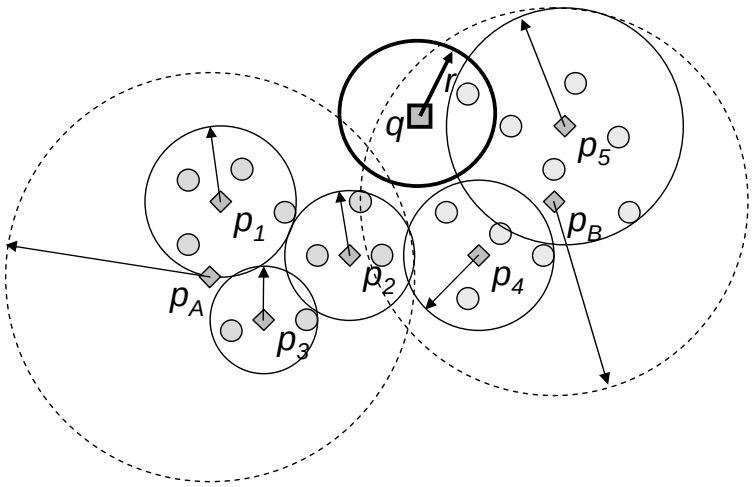
- podobný R-stromům (B^+-stromům)
- pro metrické prostory
- vyvážený, dynamický
- datové objekty uložené v listech
- interní uzly obsahují sumární informace o podstromech
- minimalizace V/V i výpočtů vzdálenosti

M-strom: struktura uzlu

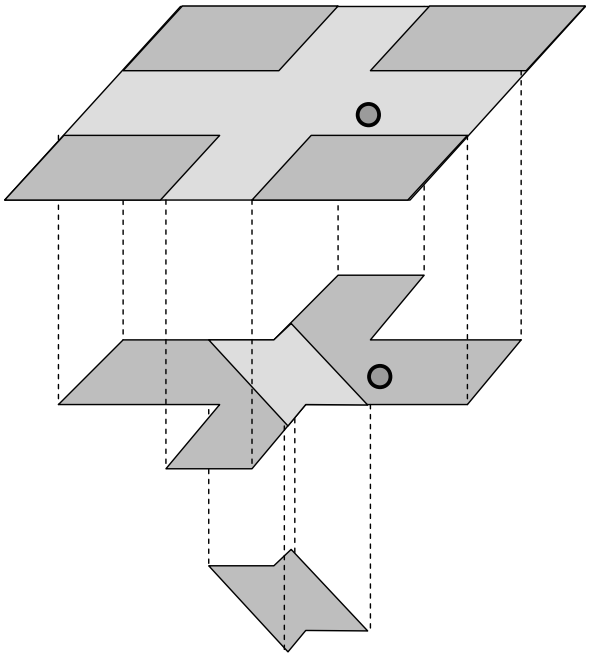
- interní uzel
  - *p* – pivot
  - *ptr(T(p))* – ukazatel na podstrom
  - *r* – krycí poloměr
  - *d(p,P(p))* – vzdálenost mezi *p* a pivotem v rodičovském uzlu
- listový uzel
  - *o* – objekt nebo jeho identifikátor
  - *d(o,P(o))* – vzdálenost mezi *o* a pivotem v rodičovském uzlu

M-strom: struktura

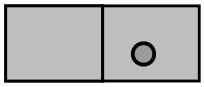




D-Index



4 oddělené kapsy (buckets) na první úrovni

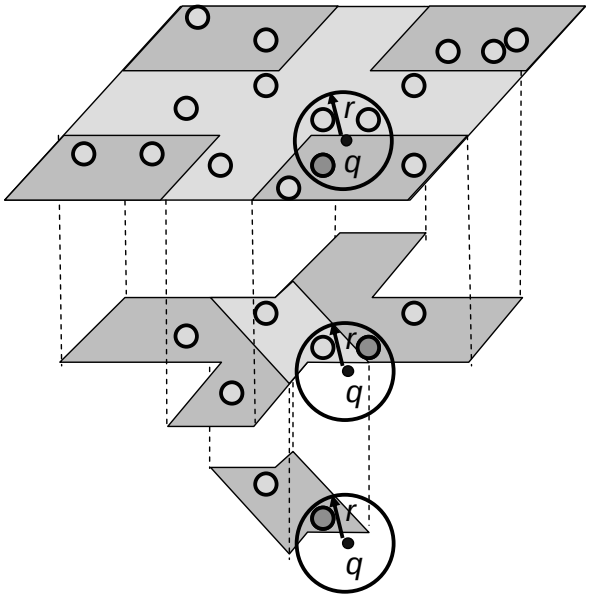


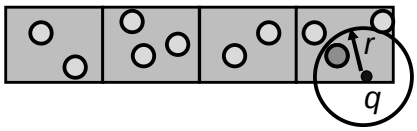
2 oddělené kapsy na druhé úrovni

Vyloučená kapsa celé struktury

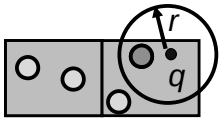
- vložení s jedním přístupem k oblasti (kapse)

Rozsahový dotaz

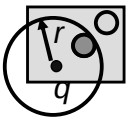




1. úroveň



2. úroveň

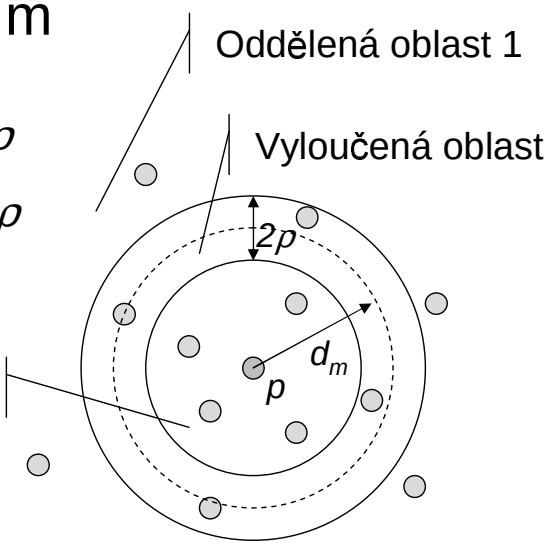


kapsa vyloučení

- **na každé úrovni přistoupena nejvýše jedna kapsa plus kapsa vyloučení**
  - **nejhorší případ (pro dotazy r ≤ ρ)**

Binární ρ-dělicí funkce

- sférické dělení s vyloučením
  - $bps(x) = \begin{cases} 0 & \text{if } d(x,p) \le d_m - \rho \\ 1 & \text{if } d(x,p) > d_m + \rho \\ - & \text{jinak} \end{cases}$

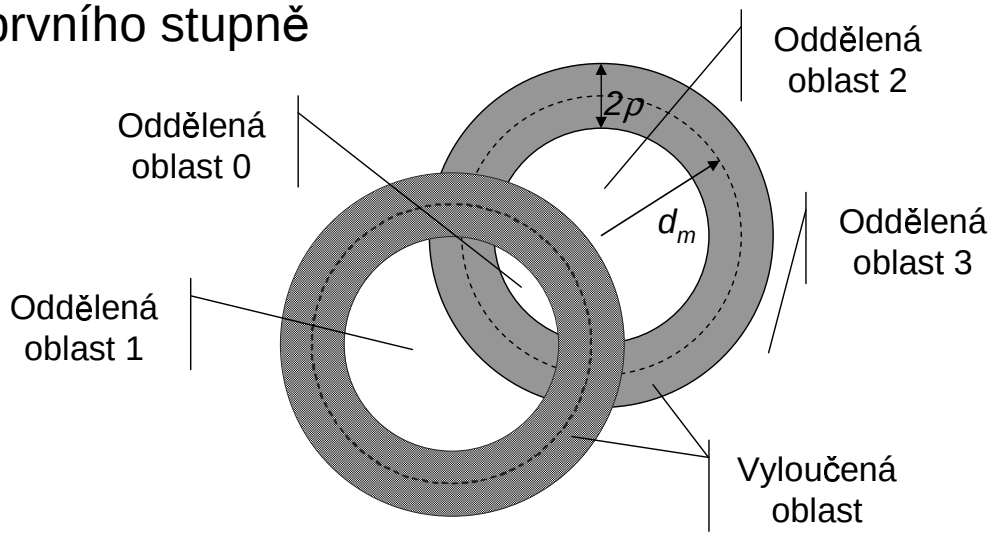


- vlastnost oddělitelnosti

  $\forall x,y \in \mathcal{D}, bps(x) = 0 \wedge bps(y) = 1 \Rightarrow d(x,y) > 2\rho$

Kombinace ρ-dělicích funkcí

- kombinace dvou ρ-dělicích funkcí prvního stupně

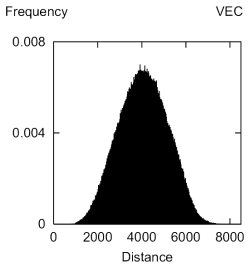


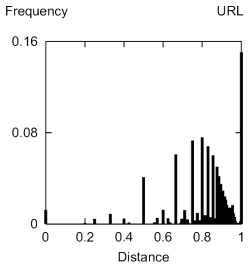
Vlastnosti D-Indexu

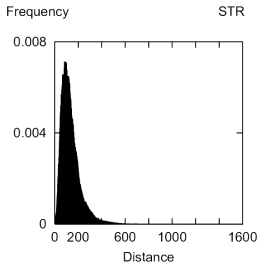
- podpora diskových úložišť
- vložení s jedním přístupem k oblasti (kapse)
  - počet volání vzdálenosti mezi m_1 a $\sum_{i=1}^{h} m_i$
- maximálně *h+1* přistoupených kapes
  - pro dotazy s poloměrem do ρ
- přesná shoda (*R(q,0)*)
  - úspěšná – přístup do jedné kapsy
  - neúspěšná – typicky bez přístupu na disk

Experimentální porovnání

- reálná testovací data
  - 45-dimenzionální vektory (quadratic form distance)
  - množiny URL adres z IS MUNI (Jacardův koeficient)
  - věty českého korpusu (editační vzdálenost)

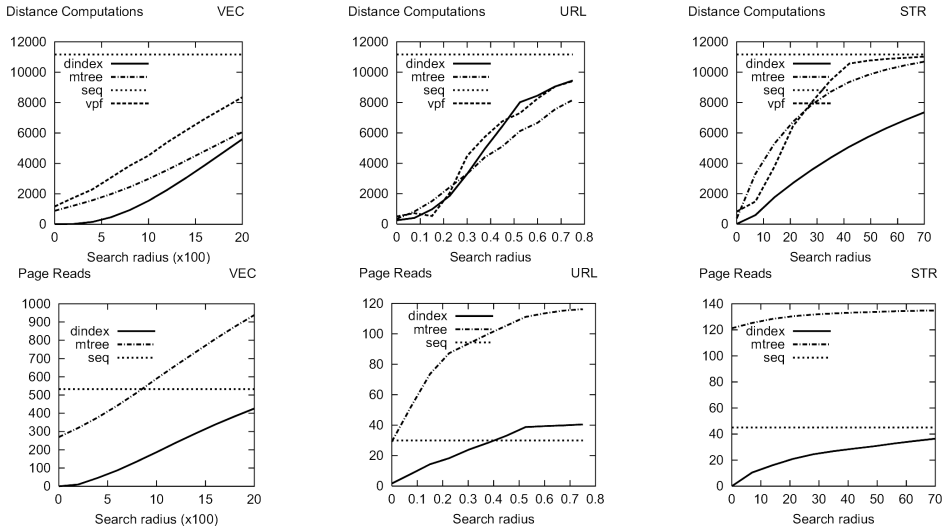






Experimentální porovnání

- rozsahový dotaz pro D-Index, M-strom, Vantage Point Forest a sekvenční přístup



Závěrem...

- distribuované indexy
  - peer-to-peer sítě
- podobnostní spojení
  - rozsahové hledání, eD-index
- vylepšení D-indexu
  - lineární hešování
- přehledový materiál
  - podobnostní hledání