



# Efektivní vyhledávání ve vektorovém modelu DIS

**Tomáš Skopal**

katedra informatiky, FEI  
VŠB-Technická univerzita Ostrava




## Osnova

- Vektorové modely DIS - problém vyhledávání:
  - v klasickém vektorovém modelu
  - v LSI modelu
- Metrické indexování vektorových modelů
  - M-strom
  - přibližné semi-metrické dotazování
- Výsledky experimentů

DIS - 25.3.2004 2

## Dokumentografické informační systémy

- Systémy pro správu a **vyhledávání** informací (zpravidla ve formě rozsáhlých kolekcí dokumentů)
- zejména textové kolekce, v poslední době multimediální databáze, XML, web, atd.
- vyhledávání v kolekci (databázi)
  - na shodu (exact-match search)
  - **podle podobnosti** (similarity search)
- modely DIS – booleovský, **vektorový**, pravděpodobnostní (a jejich modifikace)

DIS - 25.3.2004 3

## Klasický vektorový model DIS (1)

- dokument  $D_i$  modelován vektorem  $d_i$  vah termů
- mnoho způsobů konstrukce vah – nejpopulárnější *tf\*idf* (závislá na inverzní frekvenci termu v kolekci)
- kolekce reprezentována maticí termů-dokumentů
- řádká matice (max 1% hodnot nenulových)
  - používají se úsporné formáty uložení (např. CCS, CRS)
- vysoká dimenze (až stovky tisíc unikátních termů v kolekci)

DIS - 25.3.2004 4

## Klasický vektorový model DIS (2)

- funkce podobnosti (kosinová míra)  $SIM_{cos}(q, d_j)$  klasifikuje podobnost vektoru dokumentu  $d_j$  k vektoru dotazu  $q$  – kosinus odchylky vektorů

$$SIM_{cos}(q, d_j) = \frac{\sum_{k=1}^n q_k * w_{kj}}{\sqrt{(\sum_{k=1}^n q_k)^2 * (\sum_{k=1}^n w_{kj})^2}}$$

- dotazy (rozsahové a na  $k$  nejbližších sousedů) realizovány pomocí funkce podobnosti

DIS - 25.3.2004 5

## Maticе termů-dokumentů

document term \	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
database	0	0.48	0.05	0	0.70
vector	0.23	0	0.23	0	0
index	0.43	0	0	0	0
image	0	0	0.10	0	0.54
compression	0	0	0	0	0.21
multimedia	0.12	0.52	0.62	0	0
metric	0	0	0.32	0.40	0
space	0.42	0	0	0.24	0

Příklad matice termů-dokumentů

DIS - 25.3.2004 6

## LSI model

- rozšíření klasického vektorového modelu o statistické předzpracování matice termů-dokumentů (singulárním rozkladem):

$$A = U \Sigma V^T$$

- výsledkem hustá matice konceptů-dokumentů
- báze konceptů – vektory konceptů jsou lineární kombinace termů a jsou uspořádány podle důležitosti (významná témata v kolekci)
- aproximace původní matice A rank-k SVD rozkladem:

$$A \approx U_k \Sigma_k V_k^T$$

- redukce dimenze (např. ze stovek tisíc termů na stovky konceptů)
- model LSI odhaluje latentní sémantiku – díky konceptům je vyhledávání méně závislé přímo na termech  
→ LSI model (částečně) zachycuje **synonymii** a **homonymii**

DIS - 25.3.2004

7

## Matice konceptů-dokumentů

document concept \	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
concept <sub>1</sub>	-0.21	0.48	-0.05	0.10	0.70
concept <sub>2</sub>	0.23	0.20	-0.23	0.45	0
concept <sub>3</sub>	-0.43	0.02	0.32	0.24	-0.06
concept <sub>4</sub>	0.34	-0.01	0.10	0	0.54
concept <sub>5</sub>	0.31	0.9	-0.78	0.52	0.21

Příklad matice konceptů-dokumentů

DIS - 25.3.2004

8

## Problém vyhledávání ve vektorovém modelu

- matice velkého objemu (v řádu GB)
- neexistuje efektivní vyhledávací metoda

### Naivní přístup:

- sekvenční čtení vektorů dokumentů
- pro řídké vektory dotazu je efektivnější čtení vektorů termů s nenulovými váhami v dotazu
  - méně efektivní pro husté vektory dotazu (dotaz dokumentem)
  - nepoužitelné pro LSI – každý vektor dotazu je hustý

DIS - 25.3.2004

9

## Vyhledávání – čtení vektorů dokumentů

Q	document term \	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
0	database	0	0.48	0.05	0	0.70
0	vector	0.23	0	0.23	0	0
0.21	index	0.43	0	0	0	0
0	image	0	0	0.10	0	0.54
0.1	compression	0	0	0	0	0.21
0.05	multimedia	0.12	0.52	0.62	0	0
0	metric	0	0	0.32	0.40	0
0	space	0.42	0	0	0.24	0

$SIM_{cos}(d_r, d_j)$

DIS - 25.3.2004

10

## Vyhledávání – čtení vektorů termů

Q	document term \	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
0	database	0	0.48	0.05	0	0.70
0	vector	0.23	0	0.23	0	0
0.21	index	0.43	0	0	0	0
0	image	0	0	0.10	0	0.54
0.1	compression	0	0	0	0	0.21
0.05	multimedia	0.12	0.52	0.62	0	0
0	metric	0	0	0.32	0.40	0
0	space	0.42	0	0	0.24	0

$SIM_{cos}(d_r, d_j)$

DIS - 25.3.2004

11

## Metrické indexování

- vektory dokumentů = body ve vysokorozměrném prostoru
- problém je ekvivalentní vyhledávání ve vysokorozměrném prostoru na vzdálenost
- jako funkce vzdálenosti se používá **metrika d**

### Interpretace vzdálenosti jako podobnosti:

čím menší vzdálenost objektů, tím vyšší míra podobnosti dokumentů

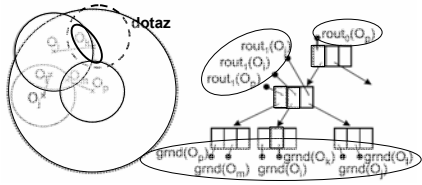
- dotaz reprezentován **regionem v metrickém prostoru** (např. rozsahový dotaz je hyper-koule definovaná středovým objektem a pokrývajícím poloměrem)
- axiom **trojúhelníkové nerovnosti** metriky dovoluje hierarchicky strukturovat celý prostor do nad- a podregionů

DIS - 25.3.2004

12

## M-strom

- datová struktura pro hierarchické metrické indexování,
- vyvážený, stránkovaný strom (ala B\*-strom, R-strom, atd.)
- **směrovací objekty** ve vnitřních uzlech představují **metrické regiony**
- v listech uloženy indexované objekty
- trojúhelníková nerovnost → možnost vyloučení nerelevantních větví stromu (resp. regionů prostoru) v průběhu dotazu



(pro 2D prostor a  $L_2$  metriku)

DIS - 25.3.2004

13

## Metrické indexování ve vektorovém modelu DIS

- realizace efektivního vyhledávání pomocí metrického indexu
- indexují se vektory (pseudo-)dokumentů, tj. sloupce matice termů-dokumentů, resp. matice konceptů-dokumentů
- z kosinové míry  $SIM_{cos}$  je odvozena **odchylová metrika** (deviation metric)

$$d_{dev}(v_i, v_j) = \arccos(SIM_{cos}(v_i, v_j))$$

- hierarchii M-stromu lze interpretovat jako hierarchii **shluků dokumentů**
- index M-stromu obsahuje pouze identifikátory objektů (sloupců) a tudíž je jeho velikost vzhledem k matici zanedbatelná (typicky do 2% velikosti matice)
- úspěšné pro LSI model (výrazné shluky – silně korelované souřadnice vektorů)

DIS - 25.3.2004

14

## Vnitřní dimenze

- pro vysokorozměrná data klasického vektorového modelu je (přesné) vyhledávání neefektivní (databázový fenomén **prokletí dimenzionality**)
  - v M-stromu se prokletí dimenzionality projevuje velkými překryvy regionů, čímž vyhledávání degeneruje na sekvenční průchod
- velikost vnitřní dimenze (**intrinsic dimension**) datové sady (jedna z definic):

$$\rho = \frac{\mu^2}{2\sigma^2}$$

kde  $\mu$  je střední hodnota a  $\sigma^2$  je rozptyl vzdáleností (na histogramu vzdáleností datové sady)

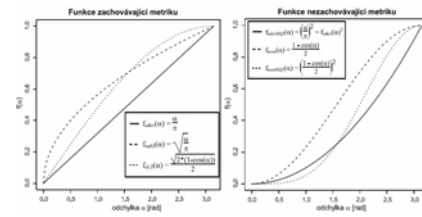
- problém vnitřní dimenze je zobrazením problému prokletí dimenzionality – datová sada s vysokou vnitřní dimenzí je **málo strukturovaná** a vyhledávání v ní je neefektivní (resp. žádná ze současných metod není efektivní)
- snahou je **snížit velikost vnitřní dimenze**, tím zvýšit vnitřní strukturovanost datové sady → efektivní vyhledávání (i za cenu pouze přibližného výsledku)

DIS - 25.3.2004

15

## Modifikace metriky (1)

- jednou z cest je použití vhodné **konvexní modifikace metriky**, ta zvýší rozptyl vzdáleností (a tím sníží velikost vnitřní dimenze)
- výsledkem je obecně semi-metrika (nesplňuje axiom trojúhelníkové nerovnosti)



konkávní modifikace

(částečně) konvexní modifikace

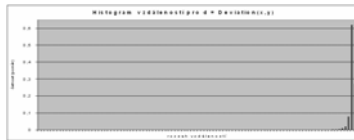
DIS - 25.3.2004

16

## Modifikace metriky (2)

Histogramy vzdáleností na vektorech dimenze 240000

a) pro původní metriku  $d$



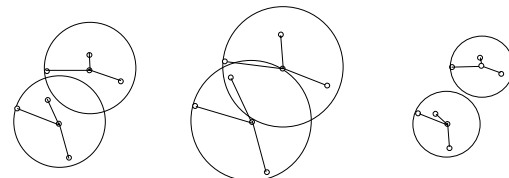
b) pro semi-metriku  $f(d)$ , kde  $f(x) = x^2$ , tj.  $f(d)$  je konvexní modifikace  $d$



DIS - 25.3.2004

17

## Modifikace metrik (3)



distribuce podle metriky  $d$

distribuce podle konkávní modifikace metriky  $d$   
**větší překryvy regionů**

distribuce podle konvexní modifikace metriky  $d$   
**menší překryvy regionů**  
(částečně narušena původní topologie)

DIS - 25.3.2004

18

## Semi-metrické dotazování

- využití konvexních modifikací metriky pro přibližné dotazování v M-stromu
- pro tvorbu M-stromu se standardně použije metrika  $d$
- modifikující funkce  $f$  se použije až při vyhodnocování dotazu, tj.  $f$  figuruje jako další parametr dotazu - rozšíření na  $range(Q, r, f)$  a  $kNN(Q, k, f)$
- měření chyby ve výsledku dotazu pomocí *relativní přesnosti* odpovědi, chyba je pro „rozumné“ funkce  $f$  malá
- výhodou je dynamická volba funkce  $f$ , a to jak pro každý dotaz, tak například i pro každou úroveň M-stromu, atd.

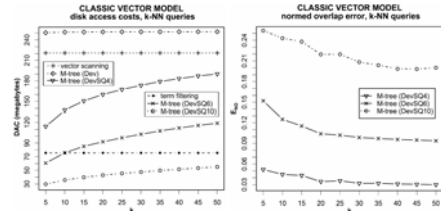
DIS - 25.3.2004

19

## Metrické indexování ve vektorovém modelu DIS, experimenty (1)

TREC LATIMES - kolekce více než 130,000 novinových článků o více než 240,000 unikátních termech

### Klasický vektorový model



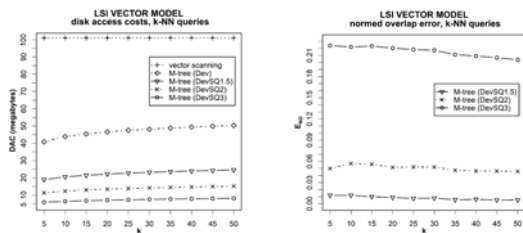
DIS - 25.3.2004

20

## Metrické indexování ve vektorovém modelu DIS, experimenty (2)

### LSI model

– původní dimenze 240000 redukovaná na 200



DIS - 25.3.2004

21

## Reference

- Skopal T., Krátký M., Snášel V.: *Efektivní implementace vektorového modelu pro dokumentografické informační systémy*, **DATAKON 2003**, Brno.
- Skopal T., Pokorný J., Krátký M., Snášel V.: *Revisiting M-tree Building Principles*. **ADBIS 2003**, LNCS 2798, Springer-Verlag, Dresden, Germany
- Skopal T., Moravec P., Krátký M., Snášel V., Pokorný J.: *An Efficient Implementation of the Vector Model in Information Retrieval*. **RCDL 2003**, St.Petersburg, Russia
- Skopal T., Krátký M., Snášel V.: *Metrické a semi-metrické indexování vektorových modelů pro dokumentografické informační systémy*, **ZNALOSTI 2004**, Brno
- 
- Skopal T., Moravec P., Pokorný J., Snášel V.: *Metric Indexing for the Vector Model in Text Retrieval*, submitted to **ACM SIGIR 2004**, Sheffield, UK
- Skopal T.: *Pivoting M-tree: A Metric Access Method for Efficient Similarity Search*, **DATESO 2004**, Desná

DIS - 25.3.2004

22