Metody dotazování pro textové databáze

Jan Martinovič

Úvod

- Vektorový model dokumentů
- Shluková analýza
- Sledování vývoje tématu
- Uspořádání odpovědi
- Algoritmus SORT-EACH
- Výsledky testů
- Závěr

Vektorový model dokumentů 1/2

- **Reprezentace dokumentu**

$$d_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,m})$$

- **Reprezentace dotazu**

$$q = (q_1, q_2, \ldots, q_m)$$

- **Indexový soubor představuje matice**
  - *i*-tý řádek odpovídá *i*-tému dokumentu
  - *j*-tý sloupce odpovídá *j*-tému termu

Vektorový model dokumentů 2/2

Indexace

Kolekce dokumentů

↓

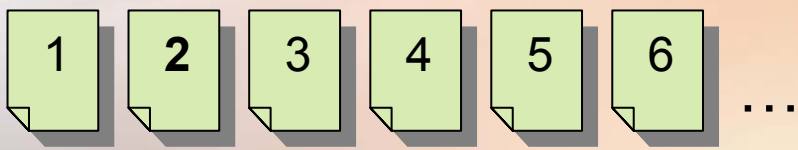
Vytvoření matice indexu

Vyhledávání

Dotaz

↓

Ohodnocený výsledek vektorového dotazu



Shluková analýza

- **Hypotéza o shlucích**

  úzce vztažené dokumenty směřují k tomu, že jsou relevantní vůči týmž požadavkům

- **Aglomerativní shlukování**

  na startu je každý dokument brán jako jeden shluk, postupně se dokumenty spojují (shlukují) dohromady.

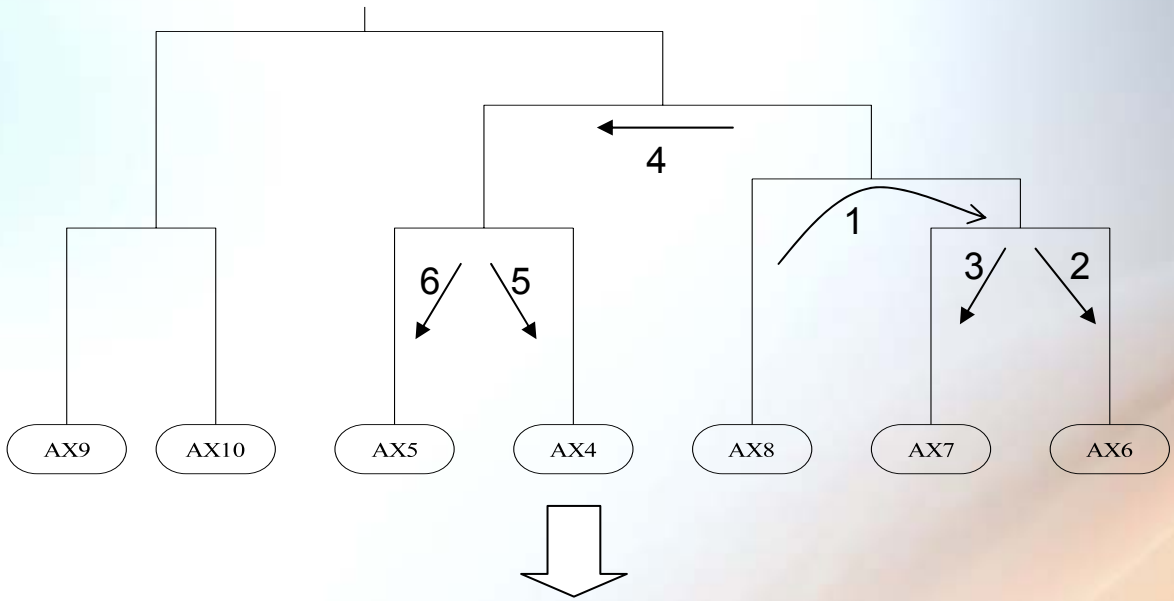
  Výpočet je ukončen v momentě, kdy všechny dokumenty tvoří jediný shluk.

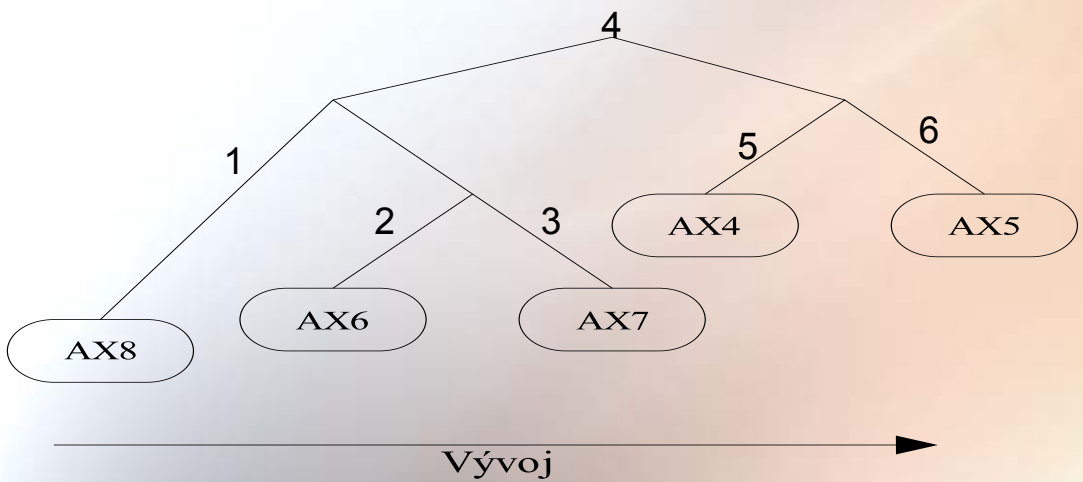
Sledování vývoje tématu 1/2

- **Cílem je k zadanému dotazu vyhledat seznam dokumentů tématicky souvisejících se zadaným dotazem.**

- **Dotazem můžou být:**
  - **Termy**
    - Je nutno nejprve vyhledat dokument nejvíce podobný zadanému dotazu.
  - **Celý dokument**
    - Procházíme hierarchii shluků od uzlu, který tvoří zadaný dokument.

Sledování vývoje tématu 2/2





Uspořádání odpovědi

- **Základní uspořádání výsledku podle koeficientu podobnosti dotazu a dokumentu**

- Algoritmus SORT-EACH provádí změnu uspořádání výsledku pomocí dotazování na vývoj tématu.

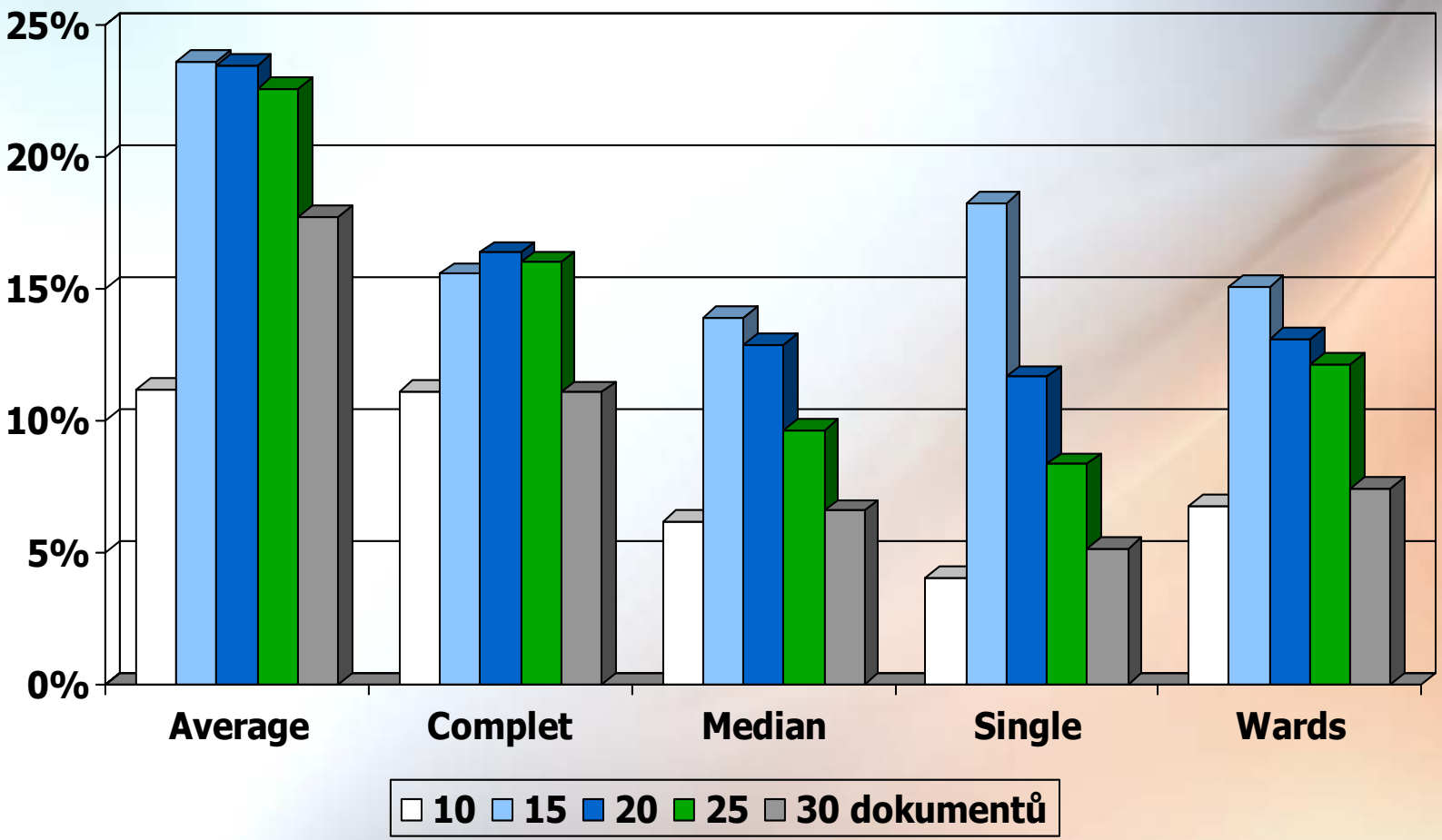
- Algoritmus SORT-EACH se snaží oddálit nerelevantní dokumenty od dotazu a přiblížit dokumenty k dotazu relevantní a tím usnadnit uživateli vyhledání požadovaných informací.

Algoritmus SORT-EACH

Testovací kolekce dokumentů

- Medlars Collection
- 1033 anglických abstraktů z oblasti medicíny
- Test proveden pro 24 dotazů
- Ke každému dotazu jsou v kolekci uvedeny relevantní dokumenty

Srovnání shlukovacích metod s metodou S-LEV6



Závěr

- Výrazné zlepšení se projevilo v prvních 10 – 30 dokumentech
- Vhodné shlukovací metody, u kterých nedochází k tvorbě řetězů
- Nejlepší se jeví hierarchické shlukování pomocí metody průměrů
- V dalším výzkumu se budu věnovat redukci časové složitosti výpočtu hierarchie shluků

Děkuji za pozornost.