# Selekční jazyky

## Současné trendy

***Přednáška č. 2 (10.3.2006)***

*Filozofická fakulta Masarykova Univerzity, Kabinet knihovnictví - Ústav české literatury a knihovnictví*

*jaro 2005/2006*

**Josef Schwarz, informační konzultant**

**http://schwarz.webpark.cz**

# Dnešní témata

- Automatická indexace
- Modely vyhledávání

# AI - vstup (přehl.studie)

- dostupnost plného textu, popř. abstraktu
- automatická/intelektuální indexace
  - AI-výhody: odstranění subjektivity
  - AI-výhody: velký objem dokumentů
  - AI-nevýhody: stroj nerozumí textu
    - Morfologie, syntaxe
    - Sémantika
      - Intratextová (Slova/výrazy, věty, odstavce, text)
      - Intertextová (různé texty)
      - Extratextová (realita)

# AI - vstup (pokr.)

- AI-problémy:
  - Pojmy nejsou vyjádřeny explicitně
  - Nepřímé odkazy na jiné části textu nebo texty
  - Text obsahuje nevýznamová slova
  - Jazykové problémy: synonymie, homonymie
  - Význam slov se mění v čase nebo mezi jednotlivými dokumenty
  - Různé tvary slov (míra závisí na jazyce)

# AI – vstup (pokr.)

- typy automatické indexace
  - extrakce (extraction indexing) – slovní indexace (**SI**)
    - klíčová slova z textu:
      - lexikální analýza (identifikace slov a sousloví)
      - odstranění nevýznamových slov
      - lematizace
      - (vážení)
      - (komparace s řízeným slovníkem)
  - přiřazování (assignment indexing) – pojmová indexace (**PI**)
    - práce s plným textem
      - pokročilé statistické a matematickolingvistické metody (pravděpodobnostní modely)
      - řízený slovník – simulace intelektuálního procesu

# SI – lexikální analýza

- Číslice
  - Odborné texty *(„§ 12")*, odborné termíny *(„MARC21")*
- Určení hranice slova
  - Mezera
  - Tečka (zkratky), spojovník *(knihovnicko-informační systém)*
  - Další interpunkční znaménka
- Velká/malá písmena

# SI – lexikální analýza (pokr.)

- Sousloví
  - Sémanticky nosnější než jednotlivá slova
  - Dvě základní metody
    - Statistická identifikace sousloví
    - Syntaktická identifikace sousloví
  - Normalizace sousloví
    - Slovník
    - Vypuštění pomocných slovních druhů a zanedbání pořadí složek
    - Syntaktická analýza s použitím kmene (kořene)

# SI – nevýznamová slova

- Odstranění nevýznamových slov
  - 20-30 % běžného textu
  - Spojky, předložky a další pomocné složky
    - Sousloví s předložkovou vazbou *(knihovny pro nevidomé)*
  - Slova bez rozlišovací funkce
- Řešení
  1. Negativní slovník (slovník nevýznamových slov, slovník stop-slov, stop-slovník)
  2. Odstranění lexikální analýzou a vážením

# SI – nevýznamová slova (pokr.)

- Tvorba stop-slovníku
  - Druhy slov (spojky, předložky, částice apod.)
  - Podle frekvence slova v textu
  - Krátká slova
    - Anti-negativní slovník

# SI – lemmatizace

- Metody
  - Algoritmické (gramatická pravidla)
    - Generování afixů
  - Slovníkově orientované
    - Slovník kmenů nebo kořenů a dalších morfologických informací
    - **Slovník afixů (sufixů a prefixů)**
  - Statistické
    - *Letter successor variety stemmer* (varieta po sobě následujících písmen)
      - Nové dokumenty v db
      - Nerozliší inflexní a derivační afixy
- Program: lemmatizátor (*stemmer*)

# SI – lemmatizace (pokr.)

- Příklady převodů slovních druhů
  - Mužský životný/ženský tvar substantiva *(autor, autorka)*, přivlastňovací přídavné jméno *(autorčin, autorův)* → mužský tvar subst., 1. pád, singulár *(autor)*
  - Adj.: stupňované tvary *(nejkonkrétnější)*, odvozená substantiva s konc. –ost *(konkrétnost)*, negace *(nekonkrétní)*, příslovce *(konkrétně)* → zákl. tvar. adj. *(konkrétní)*
  - Slovesa: časování, příč. č. a trp., slovesné jméno podstatné, opakované sloveso → infinitiv *(dělat)*

# SI – lemmatizace (pokr.)

- Lemmatizace se provádí:
  - Při indexaci
    - Malý index
    - Nutnost ručních zásahů
  - Při zpracování dotazu
    - inverzní lemmatizace (derivace)
    - Zvýšení relevance

# SI - vážení

- Různá důležitost slov pro obsah dok.
- Selektivní síla indexačního termínu (výrazu)
- Kritéria vážení:
  - Výraz (slovní druh)
  - Text (délka, počet různých termínů)
  - Vztah výrazu a textu
    - Frekvence výrazu v textu
    - Umístění výrazu ve specifické části textu (název, abstrakt, první a poslední pasáže apod.) – zohlednění koeficientem při vážení
  - Vztah termínu a celé db
    - Frekvence výrazu v db
  - Vybrané váhové funkce

# PI - vstup

- Simulace intelektuálního procesu
- Základ:
  - Výsledky SI
  - Plný text
- Předpoklad:
  - Strukturovaný řízený slovník
    - Tezarus, sémantická síť, znalostní báze

# PI - postup

- Postup PI:
  - Identifikace výrazu
  - Srovnání výrazu s relevantními profily pojmů z řízeného slovníku
  - Určení indexačních termínů
- Problémy:
  - Shoda dokument/ŘS nemusí být určující pro obsah
  - Netriviální vyjádření pojmu v textu
  - Implicitní reprezentace pojmu v textu

# AI - hodnocení

- praktické aspekty
  - plné texty
  - vyšší účinnost ve srovnání s intelektuální indexací
  - vyšší náklady – vyšší kvalita
  - oborový IS
- systémy
  - univerzální systém neexistuje
  - funkční systémy
    - specifická oblast
    - často pracují pouze s abstrakty
    - kombinace automatické a intelektuální indexace
- příklady systémů
  - ČR: (MOZAIKA), (SEMAN), KPS PČR (Parlamentní knihovna), LEGSYS
  - NASA MAI Tool (text1, text2)

# Modely vyhledávání

- booleovský model
- vektorový model
- latentní sémantické indexování (*latent semantic indexing*)

- Literatura:
  - Rauch, J. *Metody zpracování informací II. Ukládání a vyhledávání*. Praha : VŠE, 1996.
  - Pokorný, J., Snášel. V., Húsek, D. *Dokumentografické informační systémy*. Praha : Karolinum, 1998.
  - BAEZA-YATES, R., RIBEIRO-NETO, B. *Modern information retrieval*. New York : Addison-Wesley, 1999.

# Booleovský model

- teoretické základy: 50. léta 20. století
- logické operátory
  - AND, OR, NOT, XOR
    - souborný katalog AND CASLIN
    - souborný katalog OR CASLIN
    - souborný katalog NOT CASLIN
    - souborný katalog XOR CASLIN
- rozšiřování (zkracování) výrazu
  - pravostranné *(katalog*)*, levostranné *(*log)*, vnitřní rozšíření *(ka*g)*
  - rozšíření o více znaků (*), jeden znak (?)
- proximitní operátory
  - věta, odstavec, určitý počet slov (záleží/nezáleží na pořadí)

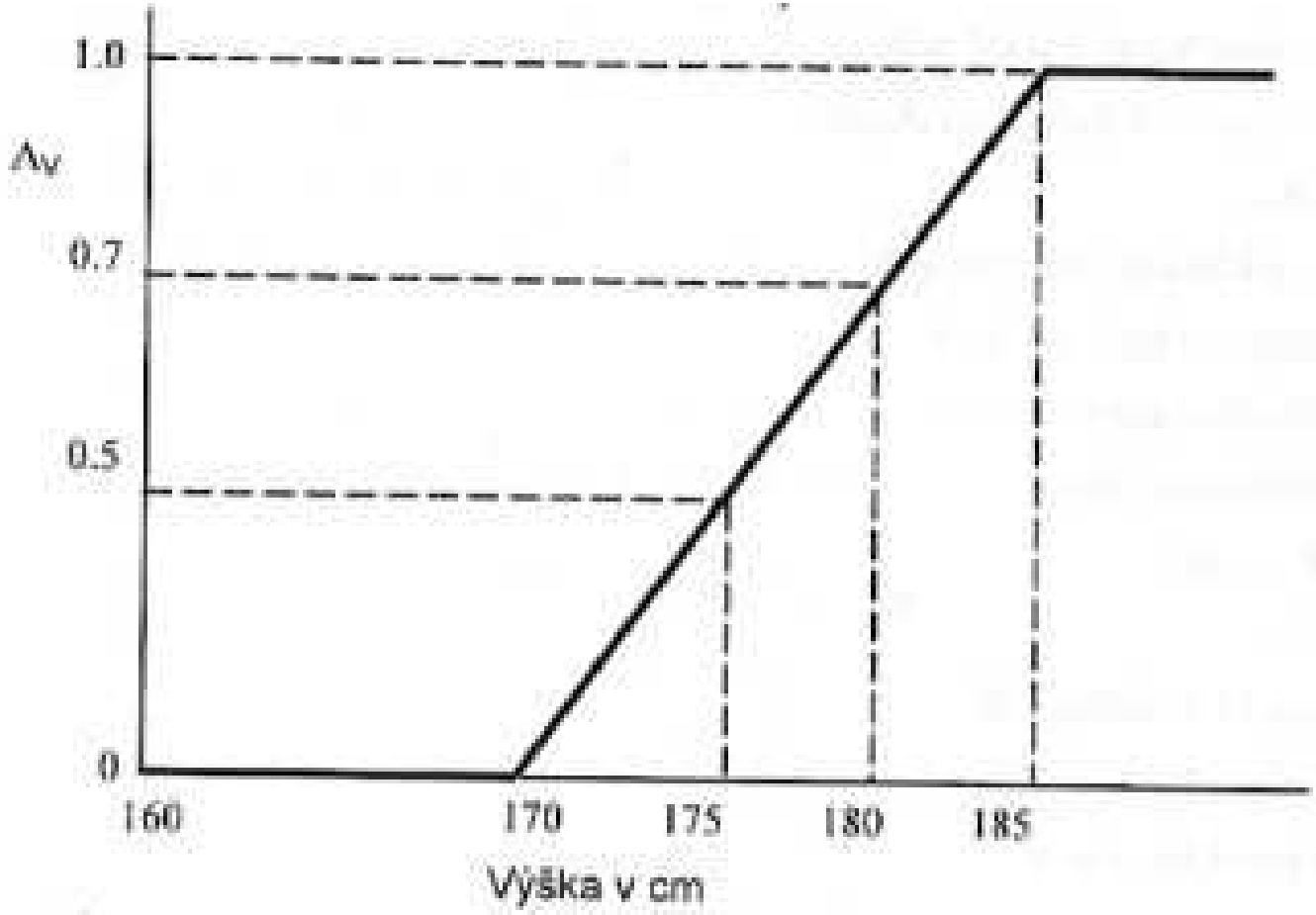
# Booleovský model

- výhody
  - jasná formalizace
  - jednoduchost
  - rychlost vyhledávání
- limitující faktory
  - úplnost, přesnost
    - použití klíčových slov
    - principiální možnosti logických spojek
      - „ostrost" – relevantní n. nerelevantní (nikoliv částečně relevantní)
      - operátor ACCRUE – systém TOPIC
    - experiment STAIRS (1985)
      - právní texty, 40 000 dokumentů
      - 51 požadavků, požadovaná úplnost: 75%
      - dosažená úplnost: 20% (přesnost 80%)

# Booleovský model

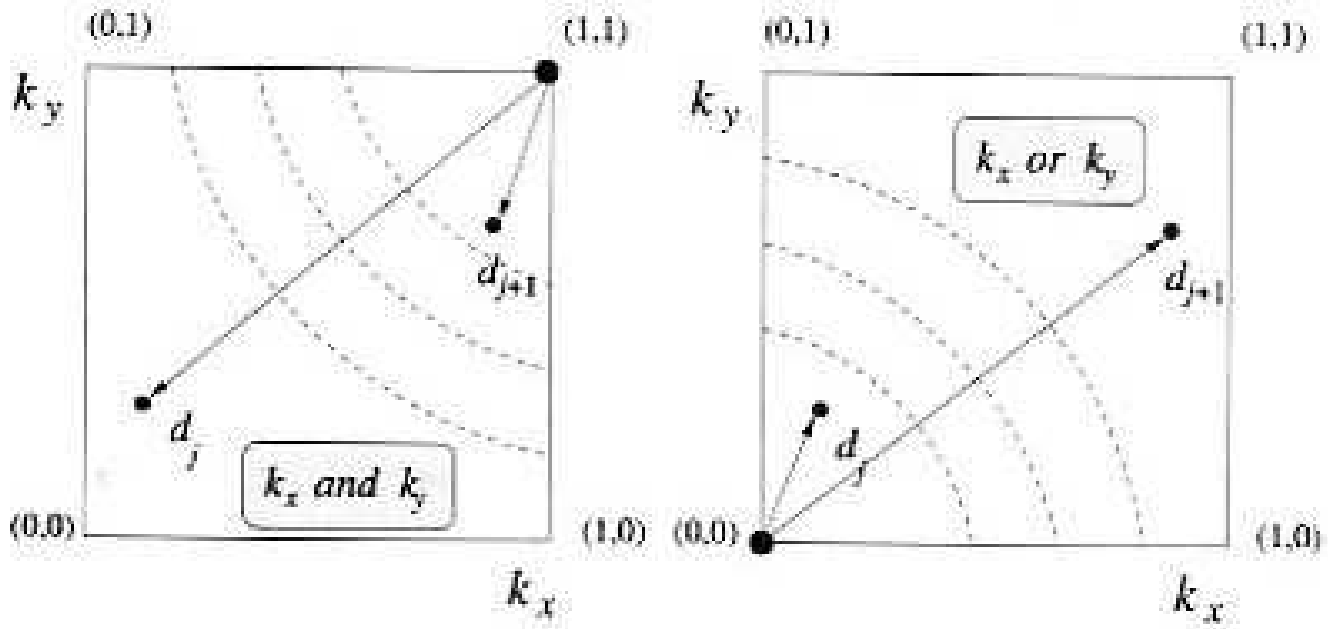
- rozšíření b. modelu
  - vážení výrazů
    - v dokumentu
    - v dotazu
  - rozšíření pomocí fuzzy logiky
    - pravdivostní hodnota z intervalu <0,1> - stupeň příslušnosti k fuzzy množině
  - geometrické rozšíření
    - dokument jako bod v prostoru
    - počet rozměrů prostoru = počet klíčových slov v dok.

# Fuzzy množina




Obr. 5.3: Spojitá funkce popisující fuzzy množinu VYSOKÝ

# Geometrické rozšíření



# Srovnání bool. mod. a rozš.

| fond | dokumentů | dotazů | přesnost pro konstantní úplnost | | |
|---|---|---|---|---|---|
| | | | booleovský model | fuzzy logika | geometrické rozšíření |
| CACM | 3 204 | 52 | 0.1789 | 0.1551 (-14%) | 0.3314 (+ 72%) |
| CISI | 1 460 | 35 | 0.1118 | 0.1000 (-11%) | 0.1806 (+ 62%) |
| INSPEC | 12 684 | 77 | 0.1159 | 0.1314 (+13%) | 0.2700 (+133%) |
| MED | 1 033 | 30 | 0.2065 | 0.2368 (+15%) | 0.5573 (+167%) |

Tabulka 8.5: Srovnání booleovského modelu a jeho rozšíření

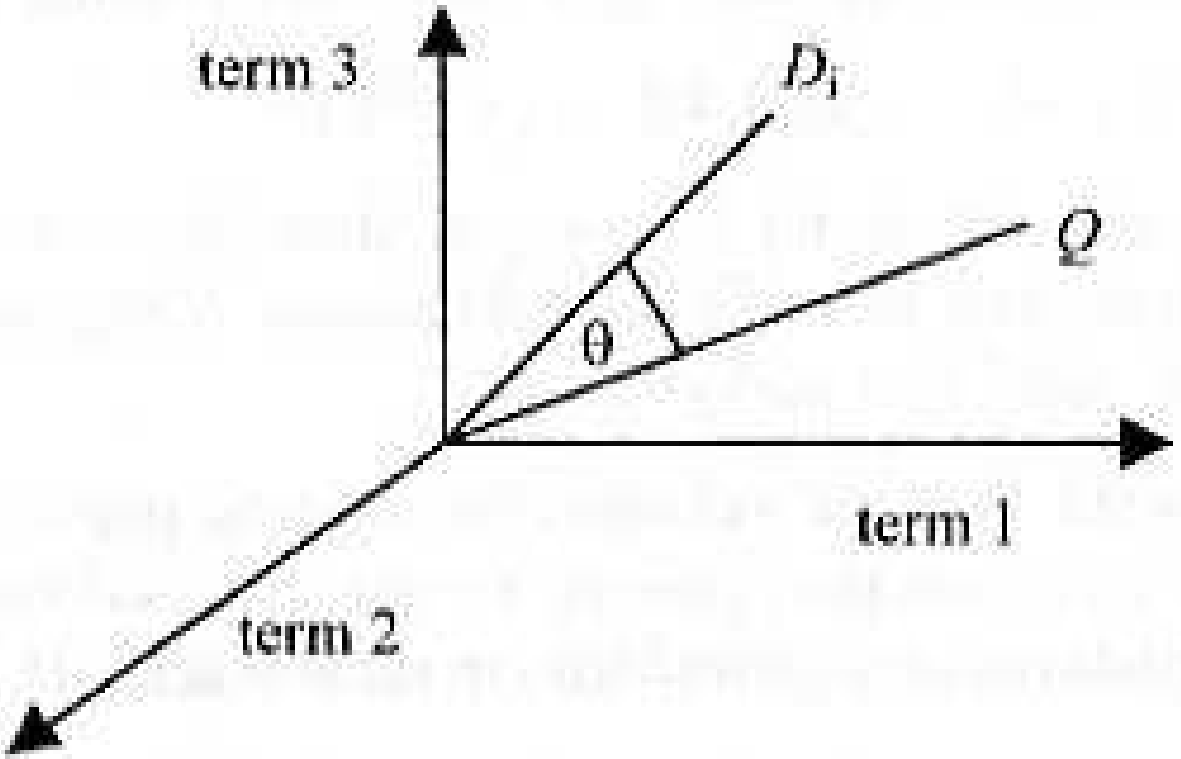
# Vektorový model

- Stupeň podobnosti mezi dotazem a dokumentem
  - vektor dotazu, vektor dokumentu
  - kosinová míra
  - váhy
- Výhody
  - vyhledává i částečně relevantní dokumenty
  - řazení dokumentů podle relevance (stupně podobnosti)
  - modifikace dotazu na základě vyhledaných relevantních dokumentů

# Vektorový model

- Nevýhody
  - není jasná interpretace vah výrazů v dotazu
  - vzorce pro měření podobnosti nejsou teoreticky zdůvodněné
  - koeficient podobnosti nemá jasný význam
  - nelze užít konjunkci a disjunkci

# Vektorový model



# Latentní sémantické indexování

- hlavní charakteristika
  - statisticko-matematické metody
  - velký objem databáze
  - základem matice dokument-výraz (klíčové slovo) → singulární dekompozice matice → redukce původní matice
  - relativně nová metoda (1988), účinnost se testuje
- Výhody:
  - pojmové vyhledávání (KS, která nebyla zadána)
  - řazení dle relevance
  - metoda nezávislá na jazyce
- příklad