Automatická indexace
Základní metody a postupy

18. 3. 2011

PŘEDMĚT: ORGANIZACE ZNALOSTÍ

PŘEDNÁŠEJÍCÍ: Josef Schwarz

AI - kontexty

- zpracování přirozeného jazyka
  - strojové zpracování textu
- AI a vyhledávání informací
- automatická klasifikace, shlukování (klastrování), abstrahování, automatická tvorba řízeného slovníku

AI - vstup

- 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

AI - příklady

- příklady systémů
  - ČR: (MOZAIKA), (SEMAN), KPS PČR (Parlamentní knihovna), LEGSYS
  - NASA MAI Tool (text1, text2)

Literatura

- Schwarz, Josef. *Současný stav a trendy automatické indexace dokumentů*. Praha, 2002-2003. Dostupné na: <http://full.nkp.cz/nkdb/docs/studie/MAIobsah.html>. Zde i další literatura.