

Současné trendy v selekčních jazycích

Přednáška č. 1 (23.2.2007)

*Filozofická fakulta Masarykova Univerzity, Kabinet knihovnictví -
Ústav české literatury a knihovnictví
jaro 2006/2007*

Josef Schwarz
schwarzjv@seznam.cz

Osnova přednášky

- ◆ Cíle
- ◆ Hlavní témata
- ◆ Dílčí témata
- ◆ Klasifikace
- ◆ Úvodní diskuse
- ◆ Automatická indexace

Cíle

- ◆ Poskytnout hlubší náhled na oblast SJ
- ◆ Prezentovat propojení oblasti SJ se souvisejícími oblastmi
- ◆ Upozornit na problémové okruhy dalšího vývoje v oblasti věcného vyhledávání informací
- ◆ Did.: rozvíjet analytické myšlení a schopnost spolupráce

Hlavní témata

- ◆ Pokročilé aplikace SJ
- ◆ Širší kontexty SJ
- ◆ Alternativní metody věcného vyhledávání
- ◆ Možnosti a limity současných SJ
- ◆ Uživatelské aplikace věcného vyhledávání

Díličí témata

- ◆ AUTOMATIZOVANÉ ZPRACOVÁNÍ DAT A SJ
 - automatická indexace, klasifikace, abstrahování, shlukování
 - automatizované zpracování přirozeného jazyka
- ◆ VYHLEDÁVACÍ MODELY A JEJICH VZTAH K SJ
 - vyhledávání, filtrace, prohlížení
 - moderní vyhledávací techniky a SJ
 - vizualizace informací
 - reprezentace a vyhledávání multimediálních dokumentů
 - citační rejstříky jako metoda věcného vyhledávání informací
 - metody zpracování a rozšiřování uživatelského dotazu pomocí řízených slovníků

Dílčí témata (pokr.)

◆ ŠIRŠÍ KONTEXTY SJ

- klasifikační výzkum, reprezentace pojmu
- formální struktura dokumentu (SGML, HTML, XML), sémantický web
- znalostní databáze, reprezentace znalostí, sémantické sítě
- ontologie a jejich vztah k SJ

◆ ŘÍZENÉ SLOVNÍKY, VĚCNÉ VYHLEDÁVÁNÍ INFORMACÍ A INTERNET

Předpoklady klasifikace

◆ Esej na zvolené (zadané) téma

- ◆ Rozsah: 15 000 znaků
- ◆ Min. počet použitých (a cit.) pramenů: 10

◆ Komparativní analýza stavu SJ/věcného vyhledávání ve vybraných informačních systémech

- ◆ Knihovny, informační střediska, souborné katalogy, oborové databáze, portály atd.
- ◆ Nedostatky systému a možná optimalizace; komparace s obdobnými systémy
- ◆ Autorský tým: 3-4 studenti
- ◆ Prezentace projektu na závěr semestru

PŘEDPOKLADY KLASIFIKACE

- dílčí nepovinné úkoly
- aktivita na přednáškách
 - ◆ průběžné získávání bodů
 - *1 bod = 1 dílek puzzle*
 - *průběžně zasílat na adresu schwarzjv@seznam.cz, předmět „puzzle“*
 - ◆ malý počet bodů = doplňkový úkol
 - studenti budou podle počtu bodů rozděleni na poloviny, dolní polovina dostane zadán úkol

Úvodní diskuse



Dnešní téma

- ◆ Automatické procedury zpracování
 - **automatická indexace**

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
 - ◆ Sousedství 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](#))