

Současné trendy v selekčních jazycích

Přednáška č. 3 (27.4.2007)

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

Josef Schwarz
schwarzjv@seznam.cz

Osnova

- ◆ Citační rejstříky jako metoda věcného vyhledávání informací
- ◆ Věcné zpracování a vyhledávání netextových dokumentů
- ◆ Sémantický web
- ◆ Sémantický grid
- ◆ Vizualizace
- ◆ Kvalita informací na internetu
- ◆ Věcné vyhledávání na webu

Citační rejstříky a věcné vyhledávání

◆ citační rejstříky

- ◆ E. Garfield
- ◆ vznik původně pro optimalizaci věcného vyhledávání

◆ kocitační analýza

- kocitace
 - 2 dokumenty citovány jedním dokumentem
 - rozsah kocitací – kocitační intenzita
- bibliografické sdružování (párování)
 - 2 různé dokumenty citují tentýž dokument

◆ skryté bibliografie

Netextové dokumenty

◆ obraz, zvuk, kombinace

- ◆ textová složka je marginální

◆ internet

- ◆ velký objem netextových informací
- ◆ omezené možnosti vyhledávání
 - [search engines](#) (podle popisku)

◆ způsoby získávání

- ◆ prohlížení
- ◆ vyhledávání

◆ Lit.: [základní přehled+další literatura](#)

Indexace netextových dok.

- ◆ podstatně složitější než indexace textových dokumentů
- ◆ hlediska indexace/vyhledávání
 - hlediska 1
 - ◆ věcnost (ofness) → „tvrdá“ indexace
 - ◆ výrazovost (aboutness) → „měkká“ indexace
 - hlediska 2
 - ◆ primitivní vlastnosti (barva, tvar)
 - ◆ logické vlastnosti (vztah mezi objekty)
 - ◆ abstraktní vlastnosti (metaforický význam)

Vyhledávání netext.dok.

◆ content-based image retrieval (CBIR)

- vyhledávání podle obsahu
- automatické zpracování obrazu (*image processing*)

◆ description-based image retrieval

- (context-based, concept-based)
- vyhledávání podle popisu (kontextu, pojmového vyjádření) (*image indexing*)

CBIR

- ◆ vyhledávání na úrovni pixelů
 - [Query by Image Content](#) (IBM)
- ◆ objektové vyhledávání
 - extrahování obrazových objektů
 - [Blobworld](#) (California University in Berkeley)
- ◆ image mining (dolování obrazových informací)
 - extrakce podobných znaků z celé db
 - ◆ [CIRES](#)
 - ◆ [SIMPLIcity](#)
 - extrakce všech vlastností bez prvotní znalosti

Vyhledávání podle popisu

- ◆ výhoda: sémantický obsah obrazu
- ◆ nevýhoda: subjektivita → inkonzistence indexace
- ◆ způsob indexace závisí na typu kolekce a požadavcích uživatelů
- ◆ indexace
 - ◆ biografických vlastností
 - ◆ předmětových vlastností
 - ◆ fyzických vlastností
 - ◆ vztahové vlastnosti

Řízené slovníky pro popis dok.

◆ ICONCLASS

- ◆ popis obrázku

◆ ATT (Art & Architecture Thesaurus)

◆ Thesaurus for Graphic Materials

- ◆ TGM I – Subject Terms
- ◆ TGM II – Genre & Physical Characteristic Terms

Aplikační oblasti

- ◆ průmyslové vlastnictví (ochranné známky)
- ◆ lékařství
- ◆ umění a architektura
- ◆ astronomie
- ◆ kriminologie
- ◆ ...atd.

Sémantický web

◆ lit.: [Sklenák, 2003](#)

◆ historie

- 2. polovina 90. let
- Berners-Lee, Hendler, Lassilla. The Semantic Web. *Scientific American*, 2001, vol. 284, May, p. 35-43

◆ základní idea

- současné způsoby vyhledávání v internetu nedostatečné
 - ◆ vyhledávače
 - ◆ portály, předmětové katalogy
- pokročilé („inteligentní“) vyhledávání v internetu
 - ◆ agenti zodpovídající komplikované dotazy
 - Který obchod prodává notebooky značky Toshiba za nejnižší cenu?
 - Kdo byl primátorem města Prahy v lednu 1946?
 - Jaká jsou aktuální rizika exportu pánských kalhot do Vietnamu?
- strukturace dokumentů
 - ◆ XML
- pojmová reprezentace
 - ◆ ontologie

◆ realizace

- pouze dílčí kroky
- fáze výzkumu – masivní nástup sémantického webu nelze v brzké době očekávat

Příklad požadavku

PC: Leo
čip: INTEL 815E
patice: Socket 370

HTML: <h1>Leo</h2>
 INTEL 815E
 <i>Socket 370</i>

XML: <PC>Kupte si naše PC</br>
 <znacka_PC>Leo</znacka_PC>
 <cip_sada>INTEL
 815E</cip_sada></br>
 <patice>Socket 370</patice>
 </PC>

Předpoklady sémantického webu

- ◆ syntaktická struktura
 - dokumenty v XML
- ◆ sémantická struktura
 - RDF – Resource Description Framework
 - ◆ objekt – atribut – hodnota
 - např. Praha je hlavní město ČR
- ◆ tvorba ontologií
 - ◆ formalizace pojmů a jejich vztahů
 - ◆ ontologie vs tezaurus
 - ◆ univerzální ontologie
 - [WordNet](#), [EuroWordNet](#)
 - ◆ doménové ontologie
 - ◆ jazyky: OWL (Ontology Web Language)

Sémantický web

Vybrané stránky

- ◆ <http://www.w3.org/2001/sw/>
- ◆ <http://www.semanticweb.org/>
- ◆ <http://www.esperonto.net>

Problémy sémantického webu

- ◆ v počátcích se předpokládal nástup sém. webu v r. 2005
- ◆ mediální bublina?
- ◆ další vývoj a výzkum?
- ◆ složitost tvorby webu pro běžného uživatele
- ◆ kontrola – „Velký bratr“?

Sémantický grid

- ◆ distribuované zpracování dat
- ◆ analýza, sdílení a výměna dat
- ◆ principy
- ◆ příklad: Medigrid

Vizualizace informací při věcném vyhledávání

◆ Kartoo

◆ Clusty

◆ Grokker

Kvalita informací přístupných prostřednictvím internetu

◆ Co je kvalita informací

- relevance (informace, které odpovídají informačnímu dotazu)
- pertinence (informace, které potřebuji)
- úplnost
- kontext (pochoitelný význam)
- spolehlivost, důvěryhodnost
- formát
- správný čas a místo

Kvalita informací přístupných prostřednictvím internetu

◆ Důvěryhodnost ([Vítů, 2005](#))

■ věrohodnost

- ◆ nezaujatost, nestrannost, objektivnost
- ◆ pravdivost
- ◆ spolehlivost, správnost, platnost
- ◆ čestnost, poctivost

■ odbornost

- ◆ zkušenost, praxe
- ◆ inteligence
- ◆ význam, vliv
- ◆ informovanost, erudice

Kritéria hodnocení

- ◆ dostupnost a použitelnost
- ◆ identifikace zdroje a dokumentace
- ◆ identifikace autora
- ◆ autorita autora
- ◆ struktura a design
- ◆ relevance a rozsah
- ◆ platnost a ověřitelnost obsahu
- ◆ přesnost a vyváženost obsahu
- ◆ navigace
- ◆ kvalita odkazů
- ◆ estetické aspekty

Kvalita informací a internet

Příklady

◆ Jiří Paroubek

- ◆ web veřejné instituce
- ◆ oficiální web veřejné osoby
- ◆ kolaborativní encyklopedie
- ◆ elektronický časopis s vyhraněným politickým postojem
- ◆ blog
- ◆ extrémně pamfletické stránky nejasného autorství

Kvalita informací a internet

Příklady

◆ teorie „volné“ energie; motionless electromagnetic generator

- ◆ „alternativní“ server
- ◆ „alternativní“ časopis
- ◆ web občanského sdružení
- ◆ osobní stránky
- ◆ web soukromé (?) „laboratoře“
- ◆ osobní stránky
-
- ◆ BIB – ceny Bludný balvan

Kvalita informací a internet

Příklady

◆ lékařské informace

■ studie

◆ validita zdrojů z hlediska lékařských informací

- 1. bibliografické a plnotextové databáze
- 2. renomované lékařské portály
- 3. ostatní dokumentu z internetu

Webové vyhledávání

Webové vyhledávání

Vybraná témata a příklady

- ◆ komparace výsledků vyhledávání prohlížečů
 - [Thumbshots Ranking](#)
- ◆ pokročilé vyhledávání
 - [Exalead](#)
- ◆ sémantické vyhledávání (sémantický web)
 - [Swoogle](#)
 - [Semantic Web Search](#)
- ◆ vyhledávání multimédií
 - [The Open Video Project](#)
 - [VideoQ](#)
 - [WebSEEk](#)

Webové vyhledávání

Vybraná témata a příklady

◆ vyhledávání v češtině

- [Morfeo](#)
- [Jyxo](#)