Věcné vyhledávání pomocí věcných SJ

Předmět: Selekční jazyky
11. 12. 2009

Přednášející: Mgr. Silvie Kořínková Presová
http://kisk.phil.muni.cz/mgr-silvie-korinkova-presova-dis

Úvod do věcného vyhledávání

věcné vyhledávání - ang. ekv. - subject searching

- tj. vyhledávání, kdy uživatel/rešeršér usiluje o nalezení dokumentů k určitému tématu X Uživatel ví, jaký dokument hledá, zná např. autora, část titulu apod.

Jeden z klíčových problémů při vyhledávání v rešeršních systémech:

Jaké vyhledávací výrazy by měly být vybrány pro formulaci dotazu?

→ *Odkud by měly být termíny vybrány?*

Úvod do věcného vyhledávání

Věcné vyhledávání lze realizovat

- pomocí pořádacích znaků věcných SJ – deskriptorů, předmětových hesel, klasifikačních znaků

- pomocí přirozeného jazyka

V praxi se doporučuje kombinovat vyhledávání pomocí přirozeného jazyka i pomocí věcného SJ – obojí v konkrétních případech přispívá ke zlepšení přesnosti a úplnosti.

Efektivní věcné vyhledávání vyžaduje následující druhy znalostí:

- Znalost polí, které mohou být pro vyhledávání využity a jejich charakteristiky.
- Znalost věcného SJ, který systém využívá.
- Znalost strategií, kde a jak je aplikovat.
- Znalost vyhledávacích možností systému a jak je použít.
- Znalost tématu.

(Poo, 2005)

Efektivní věcné vyhledávání vyžaduje následující druhy znalostí:

- Znalost toho, jak převést informační požadavek na informační dotaz.
  (Poo, 2005)

  Příklad:

- Informační požadavek:

  Využití kreativního psaní v psychoterapii či pro zlepšení duševního zdraví.

- Informační dotaz v systému Aleph MU:

  WSU = (tvůrčí psaní) AND WSU = (psychoterapie OR duševní zdraví OR psychohygiena)

Selekční jazyk - usnadňuje vyhledávání tím, že

- umožňuje kontrolovat synonyma a kvazisynonyma (*tím zvyšuje úplnost - vyhledání relevantních informací v databázi*)
  např. v hesláři databáze LLIS **Indexing vocabularies**
  Used for: Controlled vocabulary; Descriptors; Index languages; Index terms; Indexing languages; Vocabulary control
- Vyhledávání pomocí znaků SJ - stačí jeden termín **Indexing vocabularies**
- Vyhledávání pomocí neřízených termínů - k zajištění úplnosti je nutné zohlednit více variant - (index? languag?) OR (controlled vocabular?) OR (index term?)

- umožňuje rozlišit homonyma, kvalifikátor v závorce (*tím zlepšuje přesnost - vyloučení irelevantních výsledků*)
  např. Soubor věcných autorit NK ČR (SVA) *postmodernismus (literatura), postmodernismus (kultura)*

Selekční jazyk - usnadňuje vyhledávání tím, že

- poskytuje vysvětlující poznámky

např. v tezauru db LISA **Information retrieval** [+] *Very general - avoid if possible*, v SVA **Informační věda** - *Teoreticko-praktický interdisciplinární vědní obor zaměřený na výzkum a zabezpečení informačně-komunikačních procesů ve společnosti.*

- zobrazuje vztahy – hierarchické, asociace, ekvivalence – využití při specifikaci či zobecnění dotazu

např. v db LISA hledáme články o vertikálních portálech

deskriptor *Vortals*, možnost rozšířit výsledek vyhledávání pomocí nadřazeného deskriptoru *Portals*

Selekční jazyk - usnadňuje vyhledávání tím, že

- vyjadřuje termíny, které nejsou obsaženy v záznamu

např.v LLIS např Zhang Qiyu. Term selection: the key to successful indexing. The Indexer, 2009, v. 27, no. 3, s. 98-100. v SOD Indexing vocabularies

Selekční jazyk - usnadňuje vyhledávání tím, že

- odstraňuje problémy se syntaxí

Dokument je reprezentován těmito slovy v přirozeném jazyku:

např.

import, export, Česká republika, Norsko

Možné významy

- dovoz do České republiky z Norska
- dovoz do Norska z České republiky

Řešení v tezaurech – využití rolí

Řešení pomocí PH – dán kontext, hledání pomocí fráze

Selekční jazyk - usnadňuje vyhledávání tím, že

- odstraňuje problémy se syntaxí - příklad

Vyzkoušejte si vyhledat v katalogu NK ČR dokumenty týkající se dovozu do České republiky z Norska.

Vyzkoušejte si dotazy v poli Předmět:

- dotaz 1 „dovoz Česko“ „vývoz norsko“

Vyzkoušejte si vyhledat v katalogu NK ČR dokumenty týkající se vývozu z České republiky.

- dotaz 2 vývoz Česko versus dotaz 3 „vývoz Česko“

Selekční jazyk

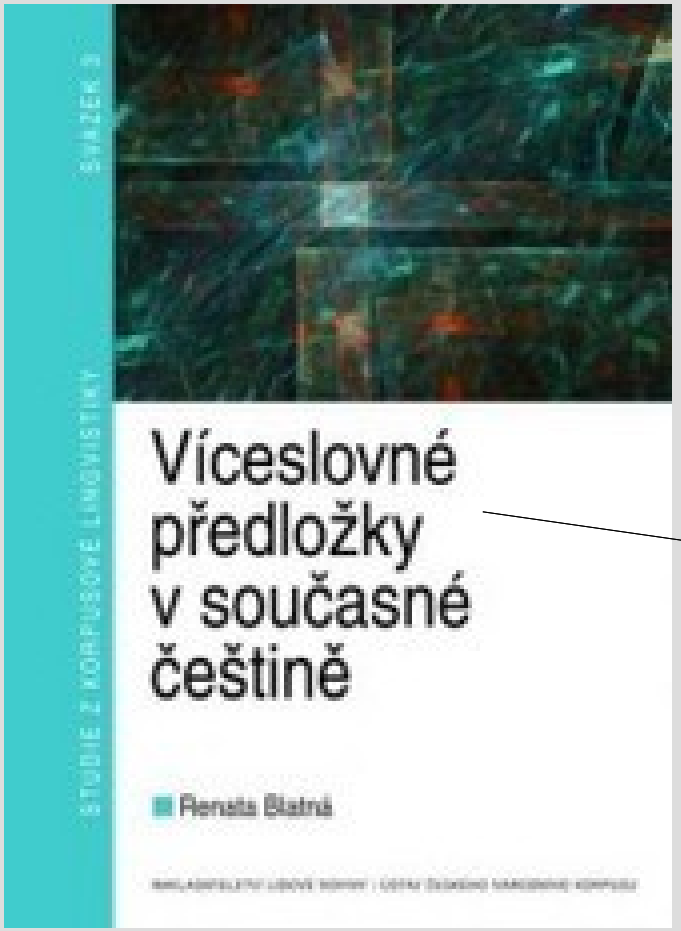
Při vyhodnocování relevantnosti výsledků vyhledávání (řazení vyhledaných záznamů) mají selekční jazyky větší váhu než slova přirozeného jazyka

PROČ?

Pořádací znak SJ byl přiřazen dokumentu na základě obsahové analýzy, z toho plyne indexace/postižení významného tématu, a to je pro vyhodnocení dotazu relevantnější.

příklad: db LLIS: http://www.hwwilson.com/Documentation/WilsonWeb/searchrules.htm

Selekční jazyk – slabé stránky



- nedostatek specifičnosti

SOD - Online katalog Národní knihovny ČR - indexace pomocí SVA

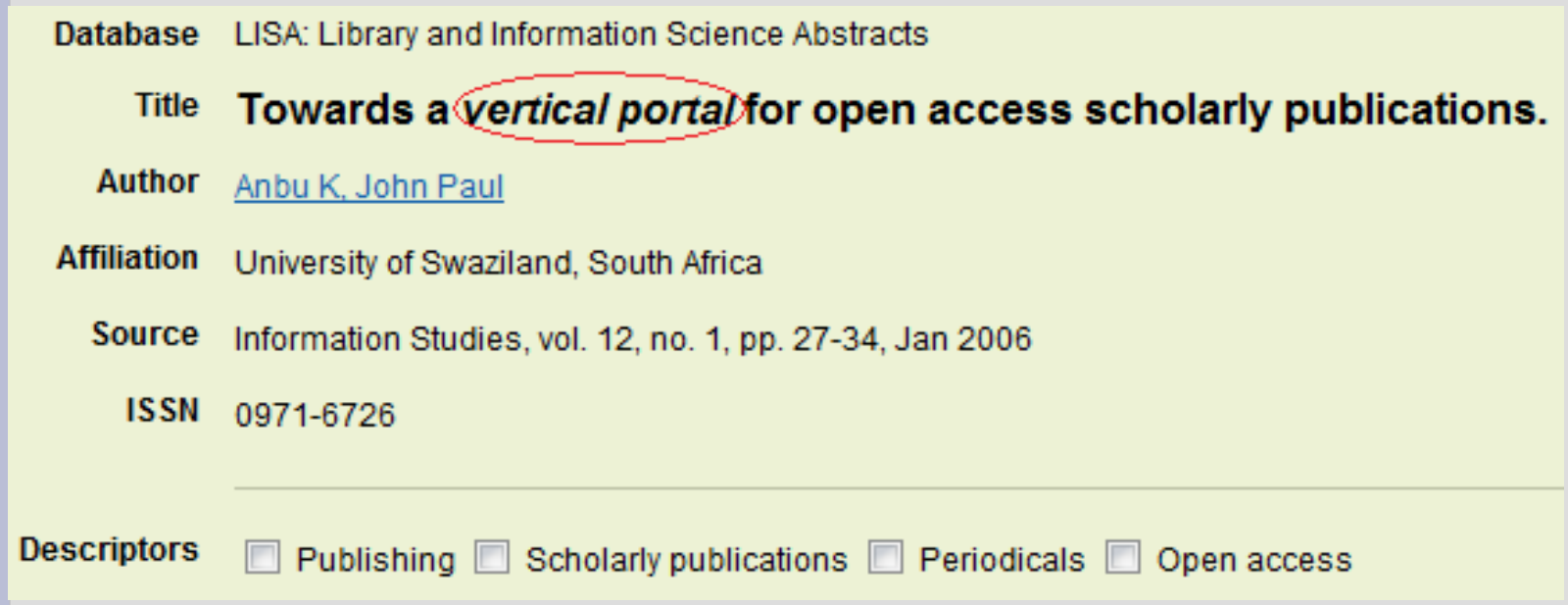
Předmět. heslo	čeština
	předložky (lingvistika)
	korpusová lingvistika
Forma, žánr	* monografie

Selekční jazyk – slabé stránky

- není okamžitá aktualizace – časová prodleva než je termín zahrnut do slovníku SJ, např. termín „ontologie“ (nástroj sémantického webu)“ v SVA
  MDT - sémantický web - Není součástí hlavních tabulek.
- slova autora mohou být nesprávně interpretovaná – nepochopení látky
- časové ztráty související s tvorbou, údržbou a osvojením si SJ

Selekční jazyk – slabé stránky

- některá témata mohou být při indexování opomenuta – např. problematika vertik. portálů v db LISA



Database	LISA: Library and Information Science Abstracts
Title	**Towards a *vertical portal* for open access scholarly publications.**
Author	Anbu K. John Paul
Affiliation	University of Swaziland, South Africa
Source	Information Studies, vol. 12, no. 1, pp. 27-34, Jan 2006
ISSN	0971-6726

Descriptors ☐ Publishing ☐ Scholarly publications ☐ Periodicals ☐ Open access

Selekční jazyk – slabé stránky

- chyby v indexaci zapříčiňují ztráty
- rešeršéři se musí učit selekční jazyk
- nekompatibilita – znesnadnění paralel. vyhledávání, bariéra snadné výměny
  - různé pořádací znaky označující jeden pojem pojem - např. označní pro *věcné SJ*

    db LLIS *Indexing vocabularies* Used for: Controlled vocabulary; Descriptors; Index languages, Index terms; Indexing languages; Vocabulary control

    db LISA *Controlled vocabulary, Index languages, Retrieval languages*
  - anglická literatura - notace **820** (DDC) X notace **PR** (LCC) X notace **821.111** (MDT)

Odlišný zkušenostní rámec indexátora a uživatele

Uživatel popisuje něco, co nezná. Na druhé straně indexátor má dokument v ruce, „všechno je před ním“.

Indexátor by měl zkoušet předvídat, podle jakých termínů budou vyhledávat uživatelé. **Jakou informaci jim daný dokument poskytne, že povede k uspokojení jejich informační potřeby?**

Odlišný zkušenostní rámec indexátora a uživatele

Indexátoři neindexují dokumenty takovým způsobem, aby zachytili nekonečně mnoho rozmanitých dotazů.

- Většinou jsou indexována hlavní a dílčí témata, tj. what is in the record.
- Nekonečně mnoho dotazů může být uspokojeno dokumentem.
- Jde o úhel pohledu - document-oriented approach x user-centered indexing

Formulace dotazu pomocí SJ

Převedení na pořádací znaky věcného SJ

→ Odvíjí se od schopnosti rešeršéra pracovat s věcným SJ (ale mnohé rešeršní systémy nabízejí řízené termíny po zadání prvního dotazu)

Převod může mít různé podoby:

1. termín v seznamu je shodný s řízeným termínem
2. termín v seznamu je synonymem/ekvivalentem – více ekvivalentů – výběr významově shodného řízeného t.
3. pro termín v seznamu existuje pouze širší termín SJ – ztráta specifičnosti původního termínu
   např. v LLIS nelze vyjádřit vertik. portály
4. pro termín v seznamu existují pouze specifičtější/podřazené termíny SJ – rozsah původního termínu je redukován
   např. v SVA – nelze vyjádřit - organizace poznání

Přirozený jazyk - výhody

- vysoká specifičnost ovlivňuje pozitivně přesnost - např. vlastní jména (osob, institucí apod.)
- schopnost vyčerpávajícím způsobem pokrýt téma, zvyšuje úplnost - neplatí u neanotovaných záznamů, zejména tam, kde je zahrnut abstrakt a plný text
- aktualizace – nové termíny jsou okamžitě dostupné
- slova užitá autorem – nemůže dojít k dezinterpretaci indexátorem
- snadnější výměna materiálu mezi databázemi – jazyková neslučitelnost odstraněna
- není třeba se jazyku učit (rodilý mluvčí)

Přirozený jazyk – slabé stránky

- intelektuální úsilí rešeršéra – problém související se synonymy (formulace dílčích dotazů) a homonymy (nutnost uvedení do kontextu)
- problémy se syntaxí – nesprávné spojení termínů, asociace – řešení pomocí proximitních operátorů
- schopnost vyčerpávajícím způsobem pokrýt téma může vést ke ztrátě přesnosti
- odlišná terminologie u jednotlivých autorů

Doporučená a použitá literatura

• Aitchison, J.*Thesaurus construction and use : a practical manual*. London : Aslib, 2000. Kapitola B1, *Is a thesaurus necessary?*, s. 5-7. ISBN 0851424465

• Bates. Indexing and Access for Digital Libraries and the Internet : Human, Database, and Domain Factors. *Journal of the American Society for Information Science and Technology*. 1998, roč. 49, č. 13.

• Chu, H. *Information representation and retrieval in the digital age*. Medford : Information Today, 2007. Kapitola 4, *Language in Information Representation and Retrieval*, s. 47-58.

• Poo, D. C. C.; Khoo, C. S. G. Online Catalog Subject Searching. In *Encyclopedia of Library and Information Science 1* [online]. 2005, č. 1 [cit. 2007-02-27]. Dostupné na World Wide Web: http://www.dekker.com/sdek/abstract~db=enc~content=a713531961

• Spink, A., et. al. Interaction in information retrieval : selection and effectiveness of search terms. *Journal of the American Society for Information Science,* 1997, roč. 48, č. 8, s. 741-61.