

Věcné vyhledávání pomocí věcných SJ - cvičení

Předmět: Selekční jazyky

10. 12. 2010

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?*

Interaction in Information Retrieval : Selection and Effectiveness of Search Terms / A. Spink, T. Saracevic

Výzkum zdrojů a efektivnosti využití vyhl. výrazů během zprostředkovaného online vyhledávání.

Identifikace 5-ti zdrojů:

- ☞ **dotaz uživatele** – termíny získané z písemně formulované žádosti, formulace informačního problému
- ☞ **interakce s uživatelem** – využití jeho znalostní struktury, termíny navržené uživatelem během interakce
- ☞ **termíny navržené řešeršérem** – před či během vyhledávání
- ☞ **řízené slovníky**
- ☞ **termíny zpětné vazby, tj. získané z vyhledaných záznamů** – termíny navržené uživatelem či řešeršérem z vyhledaných záznamů, které byly uživatelem uznány jako relevantní

Úvod do věcného vyhledávání

Věcné vyhledávání lze realizovat

- pomocí pořadacích 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 aplikací webu 2.0 v knihovnách.
 - Informační dotaz
SU(Web 2.0) AND SU(libraries)

Formulace dotazu pomocí SJ

Převedení na pořadací znaky věcného SJ

→ Odvíjí se od schopnosti řešeršera pracovat s věcným SJ (ale mnohé řešeršní 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 specifčnosti původního termínu
např. v tezauru databáze LISTA (EBSCO) nelze vyjádřit termín **thesauri** jako deskriptor, použití nadřazeného deskriptoru **subject headings**
4. pro termín v seznamu existují pouze specifčtější/podřazené termíny SJ – rozsah původního termínu je redukován
např. v SVA – nelze vyjádřit - **organizace poznání**

Formulace dotazu pomocí SJ - příklad v db ProQuest

Informační požadavek:

Využití tvůrčího psaní v psychoterapii či pro zlepšení duševního zdraví.

Pojmová analýza:

tvůrčí psaní

inventive writing

psychoterapie

psychotherapy

duševní zdraví

mental health

Deskriptory z tezauru:

Psychotherapy, Mental health, Creative writing

Informační dotaz:

SUB(Psychotherapy) OR SUB(Mental health)) AND
SUB(Creative writing)

Formulace dotazu pomocí SJ - cvičení v OPACu Aleph NK ČR

Informační požadavek: Nalezněte v katalogu NK ČR dokumenty pojednávající o postavení žen ve společnosti v ČR.

Temat. authority ze SVA

ženy

společenské poměry, společenské postavení

Česko

Informační dotaz: WKW=(česko AND ženy) AND
WKW=("společenské poměry" OR "společenské
postavení")

Formulace dotazu pomocí SJ - cvičení

- Jakými jinými tematickými autoritami ze SVA byste nahradili chybný termín **organizace poznání/pořádání informací**?
- zpracování dokumentů
- zpracování informací
- selekční jazyky
- katalogizace

Selekční jazyk - usnadňuje vyhledávání tím, že

→ umožňuje kontrolovat synonyma a kvazisynonyma (tím zvyšuje úplnost - vyhledání relevantních informací v databázi)

např. v tezauru ProQuest deskriptor Taxonomy

Used for: Scientific classification

Vyhledávání pomocí znaků SJ - stačí jeden termín/deskriptor

Taxonomy

Vyhledávání pomocí neřízených termínů - k zajištění úplnosti je

nutné zohlednit více variant - **(taxonom?) OR (scientific classification)**

→ 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 ProQuest poznámka k deskriptoru **Central content management systems** Scope Note: *Systems that enable the management of data and workflow in a collaborative environment; includes web CMS and enterprise CMS,*

v SVA **Informační věda** - *Teoreticko-praktický interdisciplinární vědní obor zaměřený na výzkum a zabezpečení informačně-komunikačních procesů ve společnosti.*

v tezauru db LISA **Information retrieval** - *Very general - avoid if possible*

Selekční jazyk - usnadňuje vyhledávání tím, že

→ **zobrazuje vztahy** – hierarchické, asociace, ekvivalence –
využití při specifikaci či zobecnění dotazu
např. v db **LISTA (EBSCO)** hledáme **články o**
folksonomiích

deskriptor *FOLKSONOMIES*, možnost rozšířit výsledek
vyhledávání pomocí asociovaného deskriptoru TAGS
(Metadata)

Selekční jazyk - usnadňuje vyhledávání tím, že

→ vyjadřuje termíny, které nejsou obsaženy v
záznamu

např. v katalogu NK ČR Hrdinové Pacifiku / Edwin
P. Hoyt

SOD:

námořní letectvo -- Spojené státy americké -- 1939-1945

námořní letci -- Spojené státy americké -- 1939-1945

stíhací jednotky -- Spojené státy americké -- 1939-1945

letecké operace -- Spojené státy americké -- 1939-1945

druhá světová válka, 1939-1945 -- Tichý oceán

Selekční jazyk - usnadňuje vyhledávání tím, že

→ odstraňuje problémy se syntaxí

Dokument je reprezentován těmito slovy
v přirozeném jazyku:

např. vyhledávací výrazy:

import, export, Česká republika, Norsko

Možné významy

☞ dovoz do České republiky z Norska

☞ dovoz do Norska z České republiky

Řešení pomocí PH – dán kontext, hledání pomocí fráze

! Vyzkoušejte v katalogu NK ČR - „**letectvo Japonsko**“ versus
letectvo AND Japonsko

Selekční jazyk

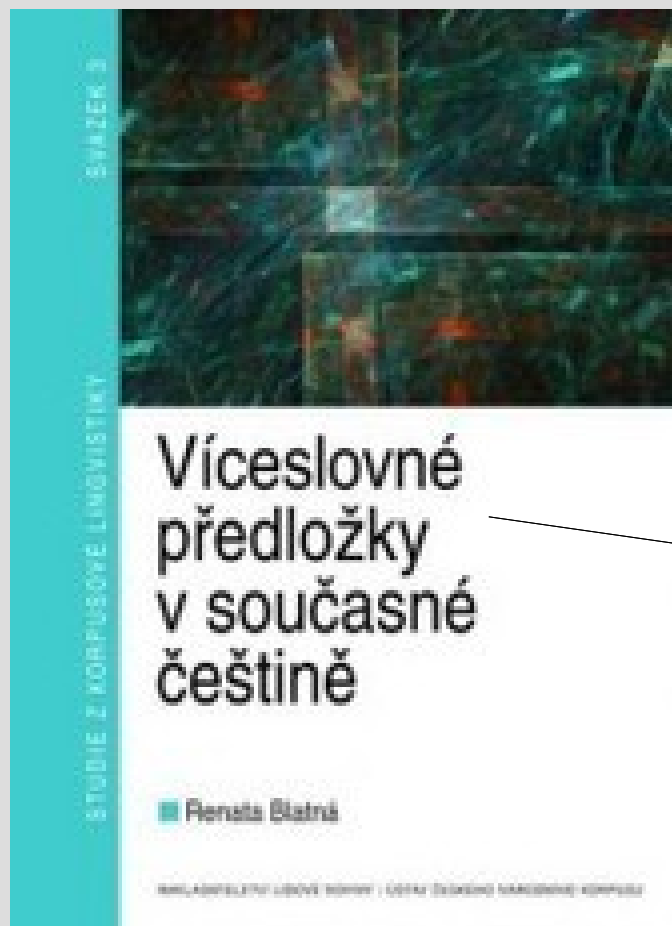
Při vyhodnocování relevantnosti výsledků vyhledávání (řazení vyhledaných záznamů) mají selekční jazyky větší váhu než slova přirozeného jazyka

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
- 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. indexace článku Kapucu, A. Getting users to library resources: A Delicious alternative. *Journal of Electronic Resources Librarianship* [serial online], December 2008; Vol. 20, Issue 4, p228-242 v DB LISTA. Chybí deskriptor FOLKSONOMIES

Selekční jazyk – slabé stránky

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

ALE

→ Nekonečně mnoho dotazů může být uspokojeno dokumentem.

→ Jde o úhel pohledu - document-oriented approach x user-centered indexing

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 řešeršéra** – 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.