Rešeršní činnost

Rešeršní strategie a věcné vyhledávání

4. 4. 2008

přednášející: Silvie Kořínková Presová
presova@phil.muni.cz
Kabinet informačních studií a knihovnictví, FF MU

Věcné vyhledávání/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?*

Výběr termínu pro formulaci dotazu a ladění rešerše

Rešeršér – dva základní okruhy zdrojů termínů:

- během interakce s uživatelem před a během vyhledávání
- během interakce s rešeršním systémem

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 rešeršérem – před či během vyhledávání
- řízené slovníky
- termíny zpětné vazby, tj. získané z vyhledaných záznamů – termíny navržené uživatelem či rešeršérem z vyhledaných záznamů, které byly uživatelem uznány jako relevantní

Věcné vyhledávání/subject searching

Dva způsoby:

- pomocí pořádacích znaků/prvků 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

Důležité termíny

- věcný SJ – umělý jazyk, „jazyk používaný pro zpracování dokumentů pomocí věcných údajů s cílem umožnit vyhledávání dokumentů podle obsahu“ (TDKIV)
  - *„Selekční jazyk je umělý jazyk určený pro vyjádření obsahu dokumentů. Skládá se z řízeného (strukturovaného) souboru lexikálních jednotek (pořádacích znaků) - řízeného slovníku, pravidel jejich tvorby a pravidel jejich užívání při věcném zpracování a vyhledávání dokumentů“* (přednáška J. Schwarz - Selekční jazyky 1, 15.10.2004)
- přirozený jazyk v IR– jazyk, kterým lidé mluví a píší, není pro potřeby IR limitován a definován (týče se slovníku, syntaxe, sémantiky, vztahů)
  - jazyk užívaný pro formulaci dotazu bez „konzultace“ řízeného slovníku

Formulace dotazu a ladění rešerše

Jde o základní okruhy využití přirozeného a selekčního jazyka.

Formulace dotazu viz přednáška č. 2 – formulace rešeršního dotaz

Ladění rešerše – query expansion (Shiri, 2002)

- manuální – uživatel se rozhodne, jak může být výsledek rešerše využit pro další úpravu dotazu
- interaktivní – uživatelé vybírají systémem navržené vyhl. výrazy (např. LLIS, ProQuest)
- automatické – vyhledané dokumenty, které označil uživatel jako relevantní jsou systémem vyhodnoceny (určení sady vyhl. výrazů pro nové hledání) a je provedeno nové vyhledávání

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
- znalost toho, jak převést informační potřebu na informační dotaz (Poo, 2005)

Selekční jazyk - usnadňuje vyhledávání tím, že

- umožňuje kontrolovat synonyma a kvazisynonyma (*tím zvyšuje úplnost - vyhledání relevantních informací v databázi*)
  např. v tezauru databáze LLIS **Indexing vocabularies**
  Used for: Controlled vocabulary; Descriptors; Index languages; Index terms; Indexing languages; Vocabulary control
- umožňuje rozlišit homonyma, kvalifikátor v závorce (*tím zlepšuje přesnost - vyloučení irelevantních výsledků*)
  např. Soubor věcných autorit NK ČR (SVA) *význam (logika)*, *postmodernismus (literatura), postmodernismus (kultura*)
  - vyzkoušejte vyhledávání v katalogu NK ČR – nejprve pomocí předmětu *postmodernismus* (zvolte vhodné pole), dále dle *postmodernismus literatura*
- poskytuje vysvětlující poznámky
  např. v tezauru db LISA **Information retrieval** [+] *Very general - avoid if possible*
  ? jaká je poznámka v tezauru ProQuest pro **Vocabularies & taxonomies**

Selekční jazyk - usnadňuje vyhledávání tím, že

- 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*
- vyjadřuje termíny, které nejsou obsaženy v záznamu

Selekční jazyk - usnadňuje vyhledávání tím, že

- odstraňuje problémy se syntaxí

Dokument je reprezentován těmito slovy v přirozeném jazyku:

- automobily, export, Spojené státy americké, Japonsko

Možné významy

- export japonských automobilů do USA
- export amerických automobilů do Japonska

Řešení v tezaurech – využití rolí

Řešení pomocí PH – dán kontext, hledání pomocí fráze

Selekční jazyk

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

PROČ?

- Termín 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 – využití při taktikách

Zúžení dotazu:

- klíčová slova se kombinují s věcným selekčním jazykem
- kombinace množiny deskriptorů/hesel s podřazenými klíčovými slovy

Rozšíření dotazu:

- dodatečné uvedení širších jednotek věcného SJ, tj. těch, které jsou nadřazeny použitým termínům (deskriptorům, předmětovým heslům) – ty naleznete v příslušných řízených slovnících
- uvedení jednotek věcného SJ jako klíčových slov (např. vyhledávání ve všech polích)

Selekční jazyk – slabé stránky

- nedostatek specifičnosti
  např. v SVA - „víceslovné předložky“
- není okamžitá aktualizace – časová prodleva než je termín zahrnut, např. termín „folksonomy“ v LISA
- některá témata mohou být při indexování opomenuta – např. problematika vertik. portálů v db LISA
  porovnejte článek ***Image Indexing : How Can I Find a Nice Pair of Italian Shoes*** v db LLIS, ProQuest
- slova autora mohou být nesprávně interpretovaná – nepochopení látky

Selekční jazyk – slabé stránky

- chyby v indexaci zapříčiňují ztráty
- rešeršéři se musí učit selekční jazyk
- nekompatibilita – znesnadnění paralel. vyhledávání, bariéra snadné výměny
  - nalezněte v tezauru db LISA deskriptor pro **Indexing vocabularies** (prefer. termín v LLIS)
  - *anglická literatura* - notace **820** (DDC) X notace **PR** (LCC)
- časové ztráty související s tvorbou, údržbou a osvojením si SJ

Odlišný zkušenostní rámec indexátora a rešeršéra/uživatele

- Uživatel popisuje něco, co nezná (zejm. první fáze viz Gaslikova, 2. přednáška). 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?**
- porozumění tématu, chápání významu slov

Odlišný zkušenostní rámec indexátora a rešeršéra/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

- více viz Bates, 1998

Formulace dotazu pomocí SJ (2. přednáška)

Převedení na termíny řízeného slovníku/věcného SJ

→ Odvíjí se od schopnosti rešeršéra pracovat s věcným SJ (ale mnohé rešeršní systémy nabízejí řízené termíny dle 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

Formulace dotazu pomocí SJ - příklady

- požadavek: články týkající se vztahu knihoven a Webu 2.0

  formulace dotazu: rešerši uskutečněte pomocí předmětového hesla/hesla z hesláře - (tj. v Subject) db LLIS
- Jakými jinými tematickými autoritami byste nahradili chybný termín organizace poznání/pořádání informací
- Jakými jinými tematickými autoritami byste nahradili chybný termín systém správy obsahu/redakční systém
- Nalezněte v katalogu MU dokumenty pojednávající o postavení žen v české společnosti (pomoci SVA)
- Nalezněte v katalogu MU dokumenty vztahující se k odívání, módě

Přirozený jazyk - výhody

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

Přirozený jazyk – slabé stránky

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

Povinná literatura

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

Doplňující literatura

- 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.
- 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
- Shiri, A. A., et. al. Thesaurus-Assisted Search Term Selection and Query Expansion : A Review of User-Centred Studies. *Knowledge Organization,* 2002, roč. 29, č. 1 (2002), s. 1-19. Dostupné též z WWW: http://eprints.cdlr.strath.ac.uk/2614/01/revie_thesaurusassisted.pdf