



VIKBA11

## **Selekční jazyky**

25. 11. 2016: Kvalita a konzistence  
indexace (P09)

FF MU, podzim 2016

Mgr. Josef Schwarz  
[126172@mail.muni.cz](mailto:126172@mail.muni.cz)



# Kvalita indexace



- kvalitní indexace – taková indexace, která zajistí (ne)vyhledání dokumentu v rámci konkrétního informačního systému
- jde o relativní hodnotu
  - ◆ účel a zaměření informačního systému
  - ◆ potřeby a požadavky uživatelů
- nelze hodnotit kvantitativními metodami
- faktory ovlivňující kvalitu indexace
- (*LAN03, kap. Quality of Indexing*)

# Konzistence indexace

- míra shody dvou nebo více SOD
- typy konzistence
  - mezi indexátory (interindexer consistency)
    - shoda indexace mezi dvěma nebo více indexátory
  - indexátora (intraindexer consistency)
    - konzistence indexace jednoho indexátora
- konzistence z hlediska hloubky indexace
  - konzistence pořadacích znaků vyjadřujících hlavní témata
  - konzistence pořadacích znaků vyjadřujících vedlejší témata
- (*LAN03, Consistency of Indexing*)

# Konzistence indexace (pokr.)

- Výpočet I.
  - dvojice indexátorů - tzv. konzistenční pár (*consistency pair*)

$$C = a / b$$

kde:

- **a** = počet pořadacích znaků, které shodně zvolili oba indexátoři
- **b** = celkový počet přidělených jedinečných pořadacích znaků

# Konzistence indexace (pokr.)

*Příklad:*

## Indexátor 1

- námořníci
- ztroskotání
- ostrovy
- Robinson Crusoe
- dobrodružné romány

## Indexátor 2

- trosečníci
- ostrovy
- dobrodružné romány
- anglická literatura
- 18. století

$$a = 2$$

$$b = 8$$

$$C = a / b = 2 / 8 = 0,25 = 25 \%$$

# Konzistence indexace (pokr.)

- Výpočet II.
  - více než dva indexátoři

**C = průměr konzistenčních párů**

# Konzistence indexace

(pokr.) *Příklad:*



Indexátor 1	Indexátor 2	Indexátor 3
<ul style="list-style-type: none"><li>námořníci</li><li>ztroskotání</li><li>ostrov</li><li>Robinson Crusoe</li><li>dobrodružné romány</li></ul>	<ul style="list-style-type: none"><li>trosečníci</li><li>ostrov</li><li>dobrodružné romány</li><li>anglická literatura</li><li>18. století</li></ul>	<ul style="list-style-type: none"><li>námořníci</li><li>ztroskotání</li><li>ostrov</li><li>Robinson Crusoe</li><li>dobrodružné romány</li><li>Pátek</li><li>anglická literatura</li></ul>

$$C_{1,2} = 2 / 8 = 0,25 = 25 \%$$

$$C_{1,3} = 5 / 7 = 0,71 = 71 \%$$

$$C_{2,3} = 3 / 9 = 0,33 = 33 \%$$

$$C = (0,25 + 0,71 + 0,33) / 3 = 0,43 = 43 \%$$

# Kvalita vs konzistence indexace



- vztah kvality a konzistence není bezprostřední
  - konzistentní indexace neznamena kvalitní indexaci
  - kvalitní indexace zahrnuje i konzistentní indexaci
  - konzistence indexace zlepšuje efektivitu vyhledávání
  - z hlediska správy databáze je konzistence kvalitou



# Faktory ovlivňující indexaci

## ○ indexátor

- obj.: zkušenost a znalost SJ, znalost oboru, porozumění textu, systematické myšlení, racionální čtení
- subj.: soustř., pečlivost, nálada, únava, motivace

## ○ SJ

- kvalita řízeného slovníku
- **indexační pravidla**

## ○ dokument/text

- obor, struktura, délka, styl, žánr, pomocný aparát, jazyk

## ○ pracovní podmínky

- prac. doba, produktivita práce, prac. prostředí, technické prostředky

# Kontrola a hodnocení indexace

## ○ v procesu indexace

### • indexační pravidla

### • srovnání s obsahem databáze

- automatické procedury: TODESCHINI, C., FARRELL, M.P. An expert system for quality control in bibliographic databases. *Journal of the American Society for Information Science*, 1989, roč. 40, č. 1, s. 1-11.

### • supervize

- správnost a úplnost obsahové analýzy
- věcná i formální správnost přiřazených pořádacích znaků
- indexační chyby

### • indexační experimenty

## ○ při vyhledávání

- relevance vyhledávání
- úplnost (*recall*) a přesnost (*precision*)

# Relevance vyhledávání

- úplnost (recall)  $R$ 
  - počet vyhledaných relevantních dokumentů / počet všech relevantních dokumentů
- přesnost (precision)  $P$ 
  - počet vyhledaných relevantních dokumentů / počet všech vyhledaných dokumentů
- poměr mezi úplností a přesností



# Indexační chyby

- data: kontrola indexace UK-ETF 1998-99
  - viz [případová studie](#)
- nejčastější typy chyb
  - opominutí hledisek (18,6%)
  - nesprávné stanovení významu nebo rozsahu deskriptoru (12,3%)
  - chybějící jednotlivé deskriptory (11,5%)
- typy chyb podle ovlivnění úplnost a přesnost vyhledávání
  - komplexní chyby snižující úplnost (23,4%)
  - dílčí chyby snižující úplnost (22,1%)
  - dílčí chyby snižující úplnost i přesnost (14,9%)

# Indexační chyby (pokr.)

typy chyb podle fáze indexace

- obsahová analýza (18,1%)

- zdroj chyb: indexátor

- identifikace pojmů (42,3%)

- zdroj chyb: indexační pravidla, indexátor

- nejčastěji: opominutí hledisek

- výběr deskriptorů z tezauru (20,8%)

- zdroj chyb: indexátor, indexační pravidla, řízený slovník



# Indexační experimenty

## 1. srovnávání různých typů selekčních jazyků

- ◆ indexace vzorku dokumentů
- ◆ porovnání formou rešeršních dotazů

## 2. konzistence indexátorů

- ◆ experimentální přístup
- ◆ dva nebo více indexátorů
- ◆ vzorek dokumentů
- ◆ indexace
- ◆ interpretace výsledků

- (případová studie)



# Využití hodnocení indexace



- indexátor

- zpětná vazba
- hodnocení práce

- SJ

- řízený slovník

- úprava lexika n. struktury
- úprava poznámek o rozsahu

- indexační pravidla

- formulace
- úprava

- dokumenty

- reindexace