



VIKMA06

# Rešeršní a studijně rozborová činnost

22. 3. 2013: Přednáška P5: Rešeršní strategie, modely vyhledávání

FF MU, jaro 2013

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



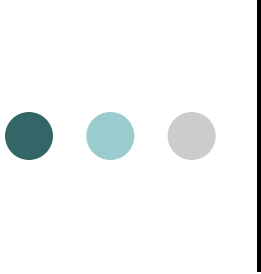
# Témata

- Formulace řešeršního dotazu
- Rešeršní strategie
- Modely vyhledávání



# Formulace rešeršního dotazu

1. pojmová analýza
2. synonyma a související pojmy
3. převedení na výrazy řízeného slovníku
4. aplikace (booleovských) operátorů
5. aplikace dalších vyhledávacích technik



# Pojmová analýza

- identifikace klíčových pojmů
- reprezentace pojmů (substantiva a adjektiva, slovesa nahrazena operátory)

- *Kdy z teplých krajin přilétají po zimě jiříčky a vlaštovky do jižních Čech?*



# Synonyma a související pojmy

- vytvoření seznamu synonym a dalších příbuzných výrazů
- využití seznamu:
  - výběru vhodného vyhledávacího výrazu
  - převod na výraz věcného SJ
  - rozšiřování a zužování tématu



# Převedení na výrazy řízeného slovníku

## Varianty

1. výraz v seznamu je shodný s výrazem ŘS
2. výraz v seznamu je synonymem/ekvivalentem výrazu ŘS
3. pro výraz v seznamu existuje pouze širší výraz ŘS
4. pro výraz v seznamu existují pouze specifičtější/podřazené výrazy ŘS



# Aplikace (booleovských) operátorů

Určení vztahů mezi pojmy

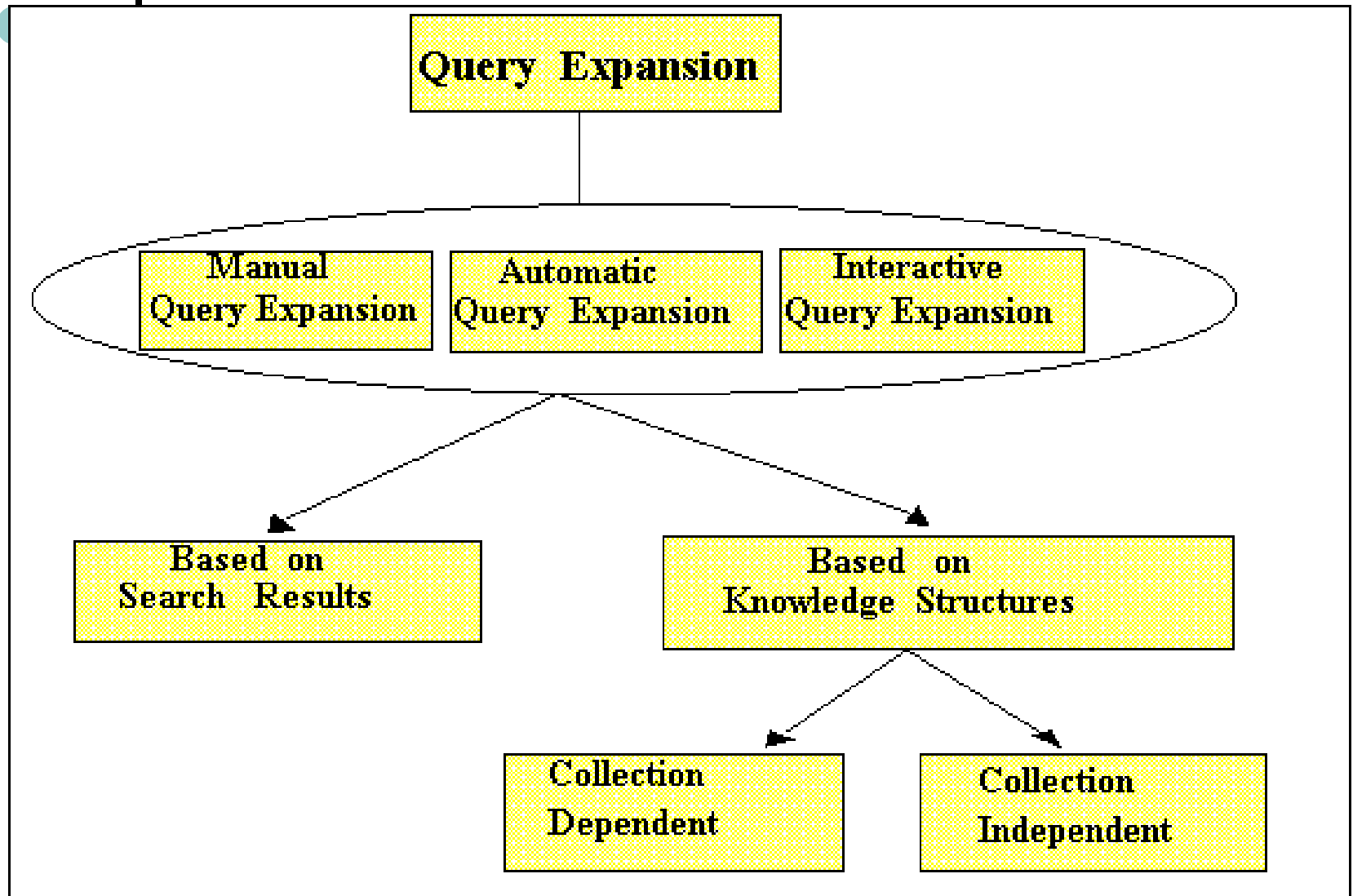
- **operátor AND** – spojení významově odlišných výrazů
- **operátor OR** – spojení synonym a příbuzných výrazů
- **operátor NOT** – vyloučení nežádoucích výrazů



# Aplikace dalších vyhledávacích technik

- škála možností závisející na konkrétním informačním zdroji
  - krácení, zástupné znaky
  - proximitní operátory
  - vyhledávání podle polí
  - rozšiřování a úprava dotazu (query expansion – relevance feedback)
  - vyhledávání ve více databázích (multiple database searching)



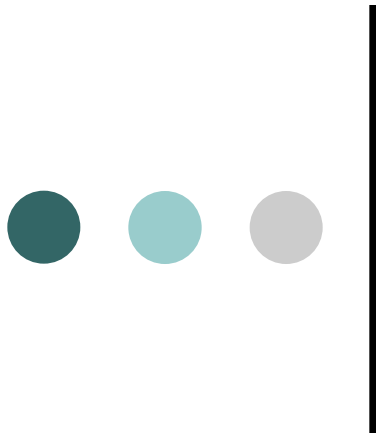




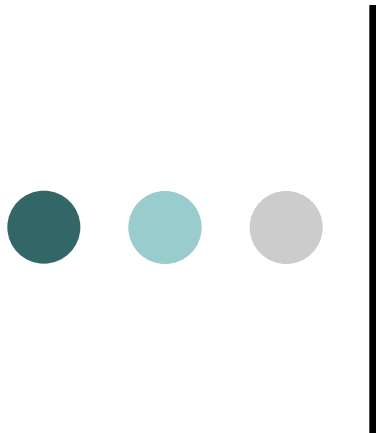
# Rešeršní strategie

- širší pojetí
- užší pojetí
  - výběr konkrétního vyhledávacího nástroje a komunikace se systémem

# Cíle řešeršní strategie

- 
- relevance X pertinence
  - úplnost X přesnost

# Úplnost a přesnost



	relevantní dok.	nerelevantní dok.
vyhledané dok.	a	b
nevychledané dok	c	d

- úplnost (R)
  - $R = a / (a+c)$
- přesnost (P)
  - $P = a / (a+b)$
- vztah mezi úplností a přesností je nepřímo úměrný (vyšší přesnost znamená nižší úplnost a naopak)

# Předpoklady řešeršní strategie

Předpokladem pro stanovení řešeršní strategie je znalost:

- informačního zdroje (databáze)
  - Obsah – jaké dokumenty, v jaké retrospektivě a úplnosti apod.
  - Struktura – podle jakých polí lze vyhledávat
- nástrojů
  - Řízené slovníky, hesláře, authority aj.
- algoritmů
  - Operátory dotazovacího jazyka, konvence pravostranného rozšíření atd.
- uživatelského rozhraní



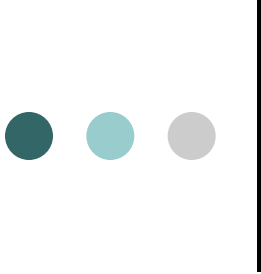
# Strategie pro zúžení záběru

- klíčová slova se kombinují s věcným selekčním jazykem
- omezení na určité pole záznamu
- využití proximitních operátorů
- omezení na určitý typ dokumentu
- operátor NOT pro vyloučení některých záznamů
- jazykové vymezení
- časové rozmezí
- kombinace množiny deskriptorů/hesel s podřazenými klíčovými slovy
- kombinace s množinou sel. údajů vyjadřující další pojem z dotazu, hledisko



# Strategie pro rozšíření záběru

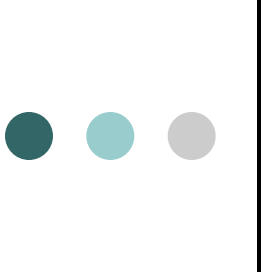
- uvedení synonym, tvarů slov, pravopisných variant (operátor OR, zástupné znaky, krácení podle slovních kořenů)
- uvedení jednotek věcného SJ jako klíčových slov (např. vyhledávání ve všech polích)
- dodatečné uvedení širších jednotek věcného SJ, tj. těch, které jsou nadřazeny použitým výrazům (deskriptorům, předmětovým heslům)
- obecné výrazy, tj. s vysokým výskytem
- zrušení předběžných omezení



# Vyhledávací techniky pro zvýšení přesnosti

- použití operátoru AND
- použití operátoru NOT
- „case sensitive“
- proximitní operátory
- vážené vyhledávání („weighted searching“)
- omezení na pole („field searching“)





# Vyhledávací techniky pro zvýšení úplnosti

- použití operátoru OR
- krácení, zástupné znaky
- fuzzy vyhledávání
- rozšiřování dotazu („query expansion“)
- paralelní vyhledávání – „multiple database searching“



# Typy řešeršní strategie

- strategie stavebních kamenů
- vyhledávání pomocí nejspecifičtější fazety
- strategie rostoucí perly
- strategie osekávání



# Strategie stavebních kamenů

- samostatné dílčí dotazy vyjadřující ústřední pojmy původního řešeršního požadavku
- identifikace klíčových/významných pojmů
- množina výrazů vztahující se k pojmu: synonyma, kvazisynonyma, pravopisné formy, nadřazené, podřízené výrazy
  - OR, truncation (krácení podle slov. kořenů), stemming, wild cards (zástupné znaky)
- spojení dílčích formulací ve finální soubor
  - AND
- vhodné použít, když usilujeme o úplnost u úzce specifikovaných témat



# Vyhledávání pomocí nejspecifičtější fazety

- ☞ vztahuje se k vyhledávání složených témat – více aspektů
- ☞ uživatel musí znát všechny dílčí témata a musí být schopen určit, které téma je nejspecifičtější
- **Vyhledávání**
  - podle nejužšího pojmu z rešeršního požadavku
  - pokud je výsledek uspokojivý, nemusí být do rešerše zahrnuta další dílčí hlediska



# Strategie rostoucí perly

Dotaz je postupně modifikován dle výsledků rešerše  
– záznamy jsou postupně procházeny a zjišťovány relevantní termíny (řízené termíny, slova z názvů apod.), které jsou použity k revidování dotazu.


**Prvotním cílem je alespoň jeden záznam**

- zjištění použitelných selekčních termínů
- úprava formulace rešeršního dotazu



# Strategie osekávání

- první formulace dotazu - **širší formulace, tj. pomocí obecného pojmu** – cílem je vyhledání více záznamů
- **postupná specifikace dotazu**
- uplatnění taktik pro zúžení záběru (AND, NOT, proxim. oper., field searching, formální omezení)
- formulace širší kategorie (obor, vědní disciplína), klasifikace
- náročnější na čas



# Rešeršní strategie - praktické rady

## Bud'te flexibilní

- berte připravené kroky strategie orientačně
- přizpůsobujte další taktiky výsledkům rešerše
- nulový výsledek – hledání příčiny

## Využívejte řízených slovníků

- využívejte souvisejících pojmů ke konkrétnímu řízenému termínu (nadřazené, podřazené pojmy)
- nikdy nespojujte výrazy s malou frekvencí výskytu (zjistitelné v katalogu) operátorem AND

## Vytvářejte množiny termínů

- je velmi důležité k jednotlivým klíčovým slovům vytvářet množiny souvisejících termínů
- termíny v množině se spojují pomocí logického součtu – OR

## Využívejte klasifikací

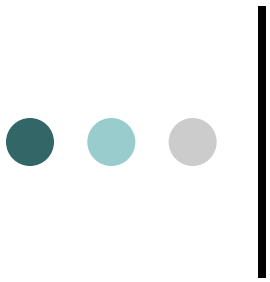
- pomocí klasifikací vyhledáte většinou mnoho záznamů, proto se hodí jejich využití při strategii osekávání



# Literatura

- kapitoly ze základní a doplňkové literatury
  - CHU07, kap. 6, 9 (s. 81-96, 145-166)
  - ING92, kap. 6 (s. 123-156)
  - VIC04, kap. 7 (s. 180-209)
- další doplňková literatura k tématu
  - Othman, R. Retrieval features for online databases : common, unique, and expected. *Online Information Review*, 2004, roč. 28, č. 3, s. 200-210.





# MODELY VYHLEDÁVÁNÍ



# Modely vyhledávání

- booleovský model
- rozšířený booleovský model
- vektorový model
- indexování latentní sémantiky (*latent semantic indexing*)



# Booleovský model

- teoretické základy (booleovská logika/algebra): 50. léta 20. století
- logické operátory
  - AND, OR, NOT, XOR
    - souborný katalog AND CASLIN
    - souborný katalog OR CASLIN
    - souborný katalog NOT CASLIN
    - souborný katalog XOR CASLIN
- rozšiřování (zkracování) výrazu
  - pravostranné (*katalog\**), levostranné (*\*komunistický*), vnitřní rozšíření (*filo?ofie*)
  - rozšíření o více znaků (\*), jeden znak (?)
- proximitní operátory
  - věta, odstavec, určitý počet slov (zaleží/nezáleží na pořadí)



# Booleovský model

## ○ výhody

- jasná formalizace
- jednoduchost
- rychlost vyhledávání

## ○ limitující faktory

### ● úplnost, přesnost

- použití klíčových slov
- principiální možnosti logických spojek
  - „ostrost“ – relevantní n. nerelevantní (nikoliv částečně relevantní)
- experiment STAIRS (1985)
  - právní texty, 40 000 dokumentů
  - 51 požadavků, požadovaná úplnost: 75%
  - dosažená úplnost: 20% (přesnost 80%)



# Booleovský model - rozšíření

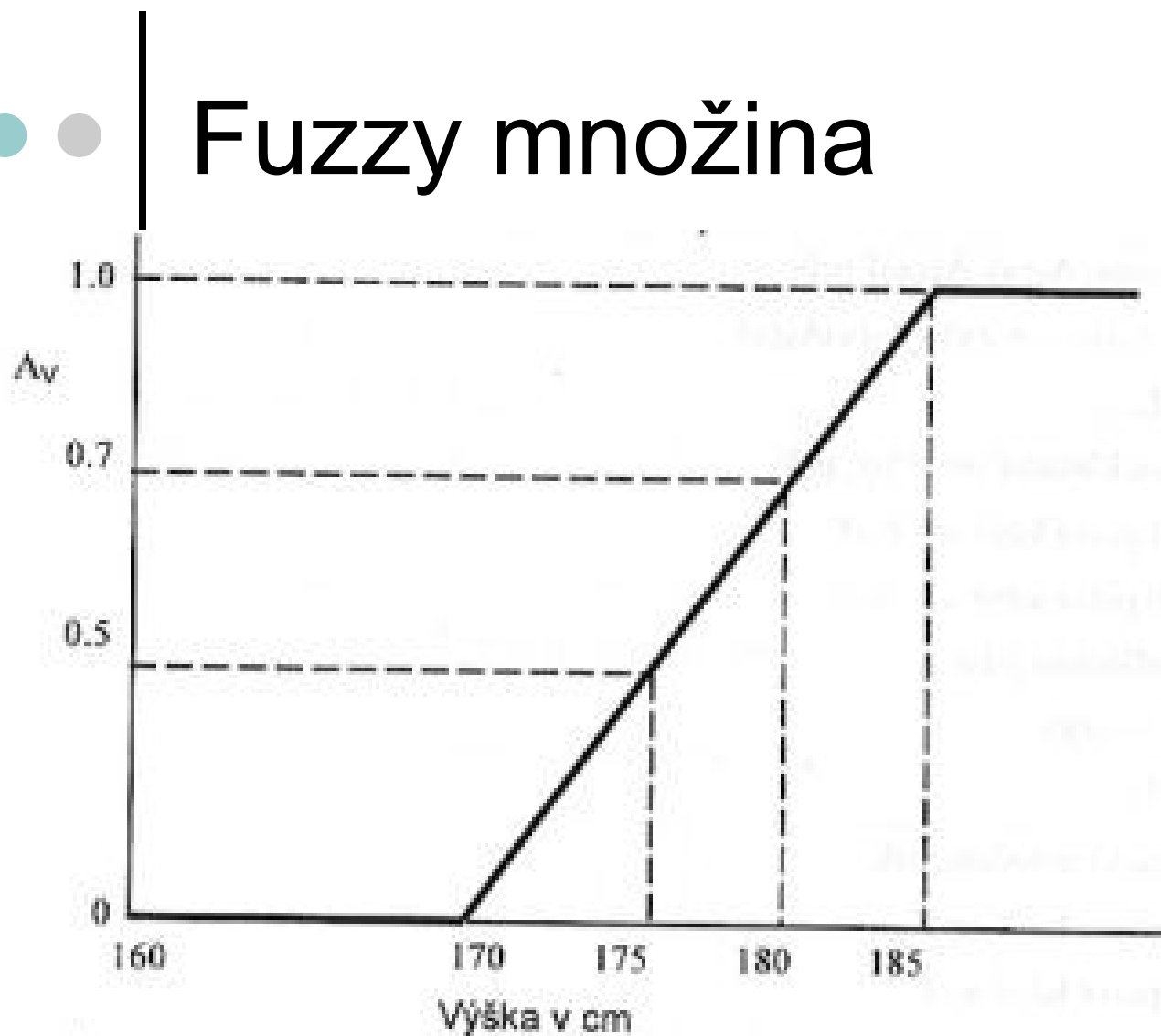
- vážení výrazů
  - v dotazu
  - v dokumentu
- rozšíření pomocí fuzzy logiky
  - formalizace principu vágnosti (schopnost přirozeného jazyka funkčně používat vágní pojmy)



# Fuzzy logika

- booleovská logika: 0/1  
(nepravda/pravda)
- fuzzy logika: pravdivost dána množinou hodnot z intervalu  $\langle 0,1 \rangle$ 
  - stupeň příslušnosti prvku do množiny

# Fuzzy množina



Obr. 5.3: Spojitá funkce popisující fuzzy množinu VYSOKÝ



# Fuzzy vyhledávání

- prvky fuzzy množiny jsou výrazy použité pro vyhledávání
- stupeň příslušnosti se určuje jako váha výrazu v dokumentu
- různé modely pro výpočet podobnosti dokumentu a dotazu

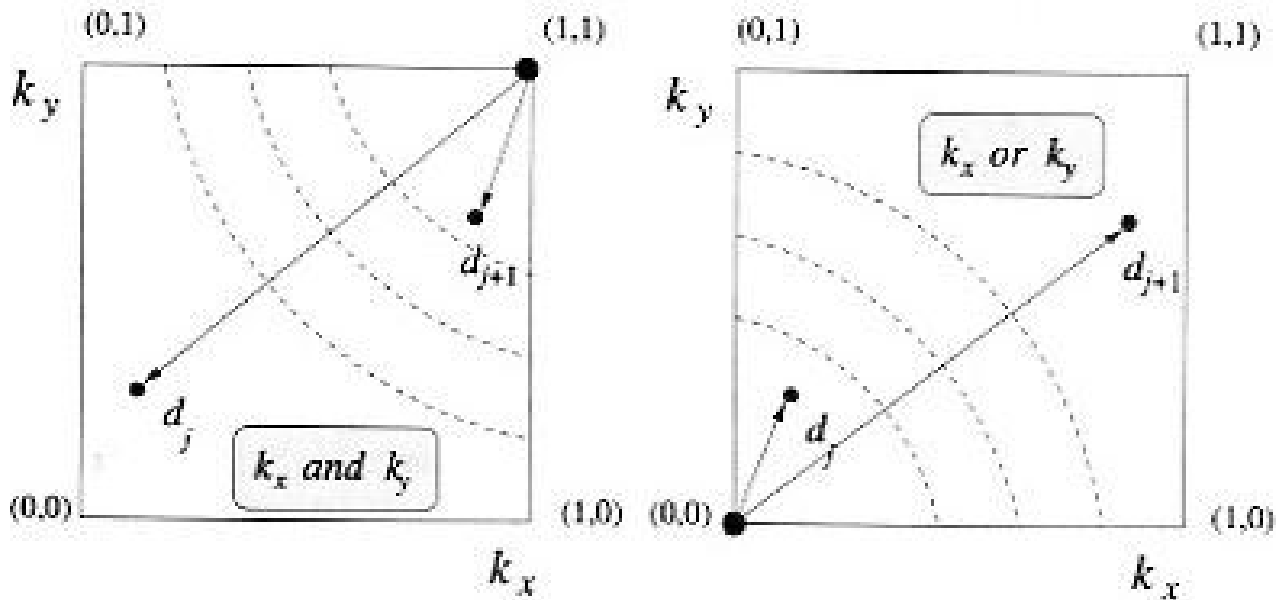




# Booleovský model - rozšíření

- geometrické rozšíření
  - dokument jako bod v prostoru
  - počet rozměrů prostoru = počet klíčových slov v dokumentu
  - vážení výrazů v dokumentu

# Geometrické rozšíření



# Srovnání booleovského modelu a jeho rozšíření

fond	dokumentů	dotazů	přesnost pro konstantní úplnost		
			booleovský model	fuzzy logika	geometrické rozšíření
CACM	3 204	52	0.1789	0.1551 (-14%)	0.3314 (+ 72%)
CISI	1 460	35	0.1118	0.1000 (-11%)	0.1806 (+ 62%)
INSPEC	12 684	77	0.1159	0.1314 (+13%)	0.2700 (+133%)
MED	1 033	30	0.2085	0.2368 (+15%)	0.5573 (+167%)

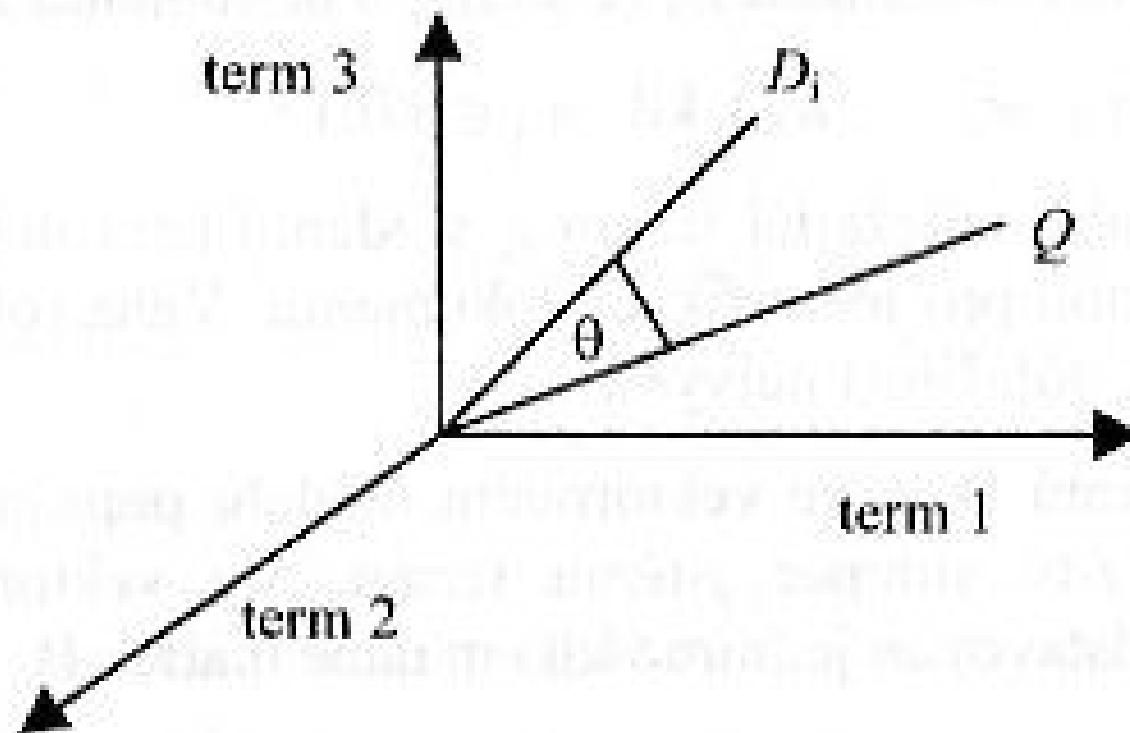
Tabulka 8.5: Srovnání booleovského modelu a jeho rozšíření



# Vektorový model

- dokument i dotaz se chápou jako vektory v  $n$ -rozměrném prostor (n je počet jedinečných výrazů ve všech dokumentech)
  - složky vektoru: směr, orientace, velikost
- složky vektorů jsou určovány výrazy a jejich vahami
- pomocí vektorového počtu se měří stupeň podobnosti mezi dotazem a dokumentem
  - kosinová míra, Diceova míra podobnosti ad.

# Vektorový model



Pokc



# Vektorový model

- Výhody
  - vyhledává i částečně relevantní dokumenty
  - řazení dokumentů podle relevance (stupně podobnosti)
  - modifikace dotazu na základě vyhledaných relevantních dokumentů



# Vektorový model

## ○ Nevýhody

- není jasná interpretace vah výrazů v dotazu
- vzorce pro měření podobnosti nejsou teoreticky zdůvodněné
- koeficient podobnosti nemá jasný význam
- nelze užít logické operátory (AND, OR, NOT)

# ● ● Indexování latentní sémantiky

- hlavní charakteristika
  - statisticko-matematické metody
  - velký objem databáze
  - základem matice dokument-výraz (klíčové slovo) → singulární dekompozice matice (redukce původní matice) → matice pojem-pseudodokument (odhalení vztahu mezi souvisejícími výrazy a zjištění podobných dokumentů)
- Výhody:
  - pojmové vyhledávání (vyhledají se i dokument obsahující výrazy, která nebyly zadány do dotazu, ale přitom jsou sémanticky blízké)
  - řazení dle relevance
  - metoda nezávislá na jazyce
- Nevýhody:
  - výpočetní náročnost





# Literatura

- kapitoly ze základní a doplňkové literatury
  - CHU07, kap. 4 až 5, 7 (s. 47-80, 97-116)
  - RAU96, kap. 6 až 10 (s. 33-57)
  - ING92, kap. 4 (s. 61-81)
  - BAE99, kap. 2 (s. 19-71)
- další doplňková literatura k tématu
  - Pokorný, J., Snášel, V., Húsek, D. *Dokumentografické informační systémy*. Praha : Karolinum, 1998, kap. 5 (s. 83-113)