VIKMA06

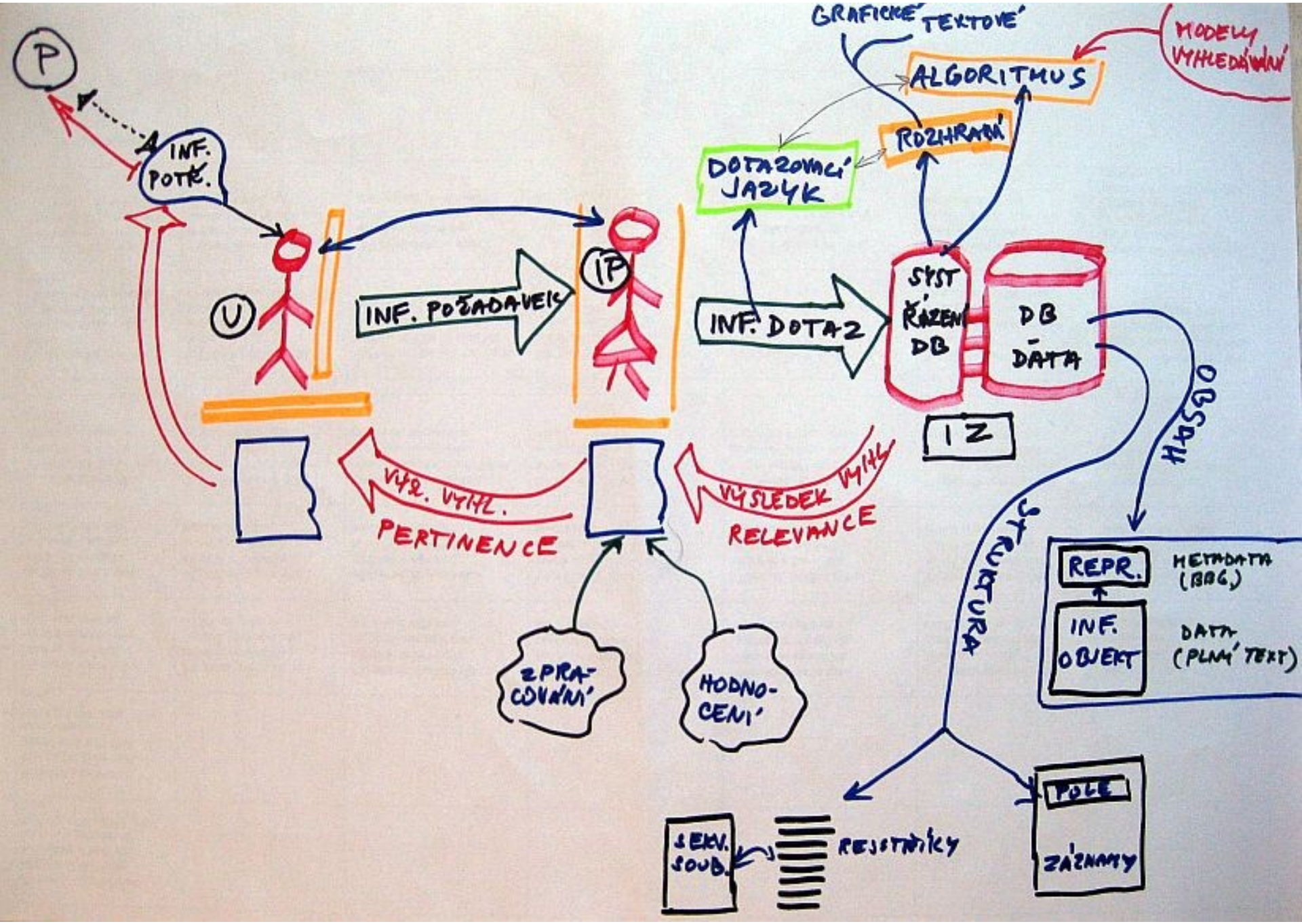
# Vyhledávání informací

14. 10. 2016: Přednáška P4+K2: Metodika vyhledávání, modely vyhledávání

FF MU, podzim 2016

Mgr. Josef Schwarz
126172@mail.muni.cz

# STRUČNÝ SOUHRN DOSAVADNÍ LÁTKY


P
INF. POTŘ.
U
INF. POŽADAVEK
IP
INF. DOTAZ
DOTAZOVACÍ JAZYK
GRAFICKÉ
TEXTOVÉ
ALGORITMUS
ROZHRANÍ
MODELY VYHLEDÁVÁNÍ
SYST. ŘÍZENÍ DB
DB DATA
IZ
VÝSLEDEK VYHL.
RELEVANCE
VÝS. VYHL.
PERTINENCE
ZPRACOVÁNÍ
HODNOCENÍ
HASOB
STRUKTURA
REPR.
METADATA (BAG)
INF. OBJEKT
DATA (PLNÝ TEXT)
SEKV. SOUB.
REJSTŘÍKY
POLE
ZÁZNAMY

# Metodika vyhledávání

- typy vyhledávání
- nástroje vyhledávání
- formulace rešeršního dotazu

# Typy vyhledávání

- Podle hledané informace:
  - identifikační vyhledávání
    - (známe některé údaje o hledaném dokumentu nebo položce)
    - vyhledávací výrazy: formální údaje - osobní jméno, název, nakladatel, rok, místo vydání, název časopisu, ISBN, ISSN, datum (konání konference, vydání, narození aj.) apod.
    - příklad: NTK, telefonní seznam, Obchodní rejstřík
  - věcné vyhledávání
    - (neznáme požadovaný dokument, hledáme určité téma)
    - vyhledávací výrazy: věcné údaje - klíčová slova z názvu, předmětová hesla, klíčová slova, deskriptory tezauru, klíčová slova z textu dokumentu (redukovaného nebo plného), klasifikace (MDT, OKEČ, NAICS) apod.
  - faktografické
    - (chceme zjistit konkrétní informaci)
    - vyhledávací výrazy: údaje podle obsahu a struktury zdroje

# Typy vyhledávání

- hledání (seeking)
- vyhledávání (searching)
- prohlížení (browsing)
- filtrace (filtering)
- data mining

# Typy vyhledávání

- nestrukturované (freetextové)
  - celý záznam dokumentu
- strukturované
  - metadata
    - selekční obraz dokumentu
    - redukovaný text
  - vazby dokumentů
    - citační vazby
    - formální vazby (FRBR)
- plnotextové

# Typy vyhledávání

- nestrukturované vyhledávání
  - základní, jednoduché vyhledávání
    - KNAV, PubMed, Google
- strukturované vyhledávání
  - pokročilé, podrobné vyhledávání
    - KNAV, PubMed, Google
  - řízený slovník (tezaurus, seznam předmětových hesel nebo klíčových slov apod.)
    - není dostupný: KNAV
    - je dostupný samostatně: NTK
    - je dostupný při vyhledávání: NKP
- plnotextové (fulltextové) vyhledávání
  - invertovaný rejstřík
  - sekvenční vyhledávání

# Nástroje vyhledávání

- vyhledávací (dotazovací) jazyky
  - složky
  - standardizace (CCL)
  - tendence ke (kvazi)přirozenému jazyku
- selekční jazyky
  - věcné
  - identifikační (autority)
  - sémantické sítě
- uživatelské rozhraní
  - příkazový řádek
  - GUI
- algoritmy 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)

# 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

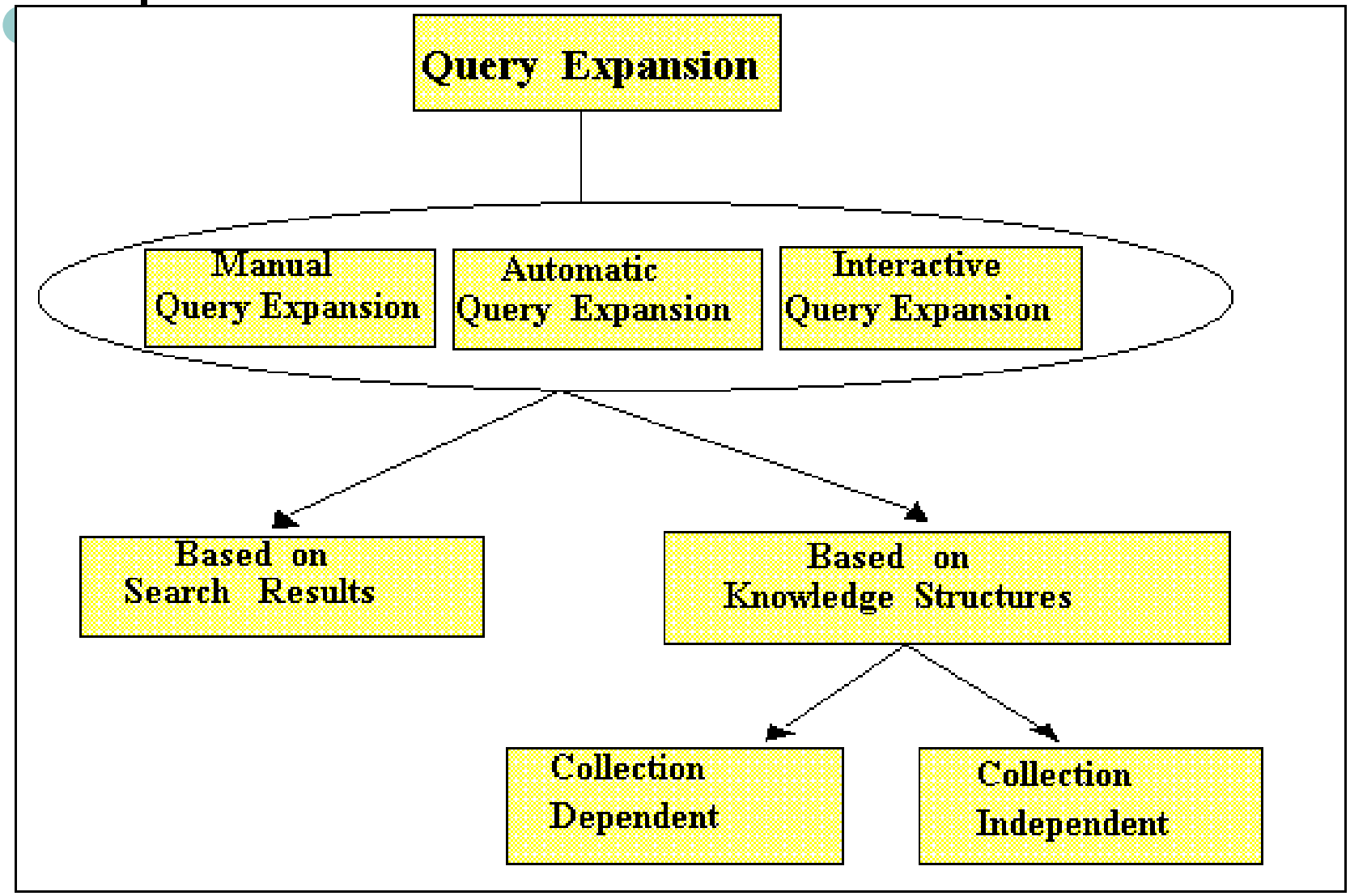
# Aplikace (booleovských) operátorů

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)


Query Expansion
Manual Query Expansion
Automatic Query Expansion
Interactive Query Expansion
Based on Search Results
Based on Knowledge Structures
Collection Dependent
Collection Independent

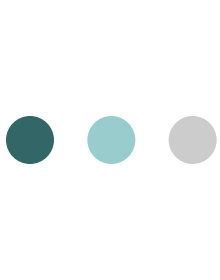
# Vyhledávací techniky

## obvyklé možnosti

- booleovské operátory
- fráze
- vyhledávání podle polí
- formální omezení
- krácení, zást. znaky, stemming
- ukládání rešerše a historie
- proximitní vyhledávání
- užití řízených slovníků

## specifické možnosti

- prohlížení časopisů a obsahů jednotlivých titulů
- rozšiřování dotazu
- navrhování výrazů ŘS
- dotaz příkladem
- automatický překlad
- odkazy na plný text prostřednictvím jiné služby, odkazy na web, napojení na katalog
- vyhledávání pomocí notací SSJ



# MODELY (TECHNIKY) 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)
      - operátor ACCRUE – systém TOPIC (příklad + příklad aplikace)
    - experiment STAIRS (1985)
      - právní texty, 40 000 dokumentů
      - 51 požadavků, požadovaná úplnost: 75%
      - dosažená úplnost: 20% (přesnost 80%)

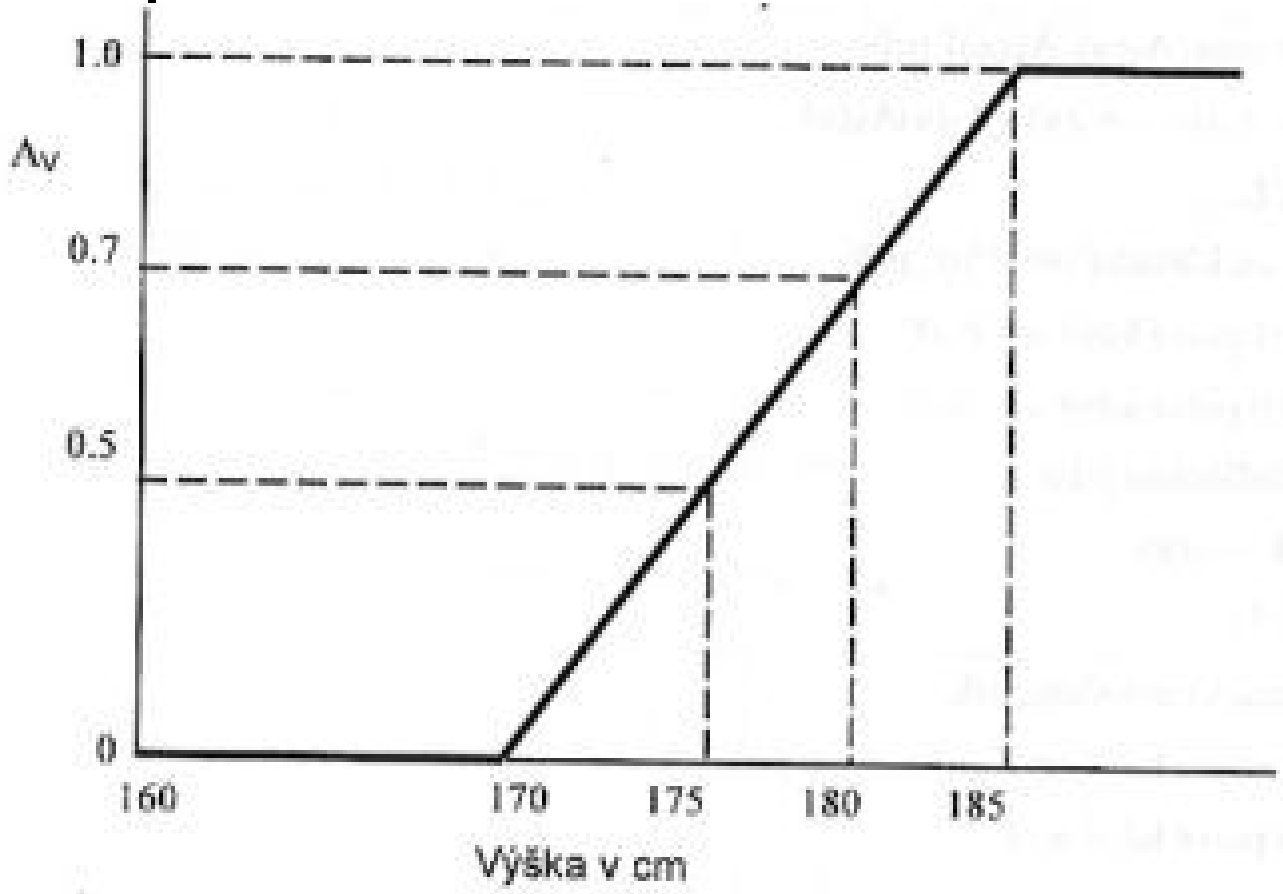
# Booleovský model - rozšíření

- 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 <0,1>
  - stupeň příslušnosti prvku do množiny

# Fuzzy množina




Obr. 5.3: Spojitá funkce popisující fuzzy množinu VYSOKÝ

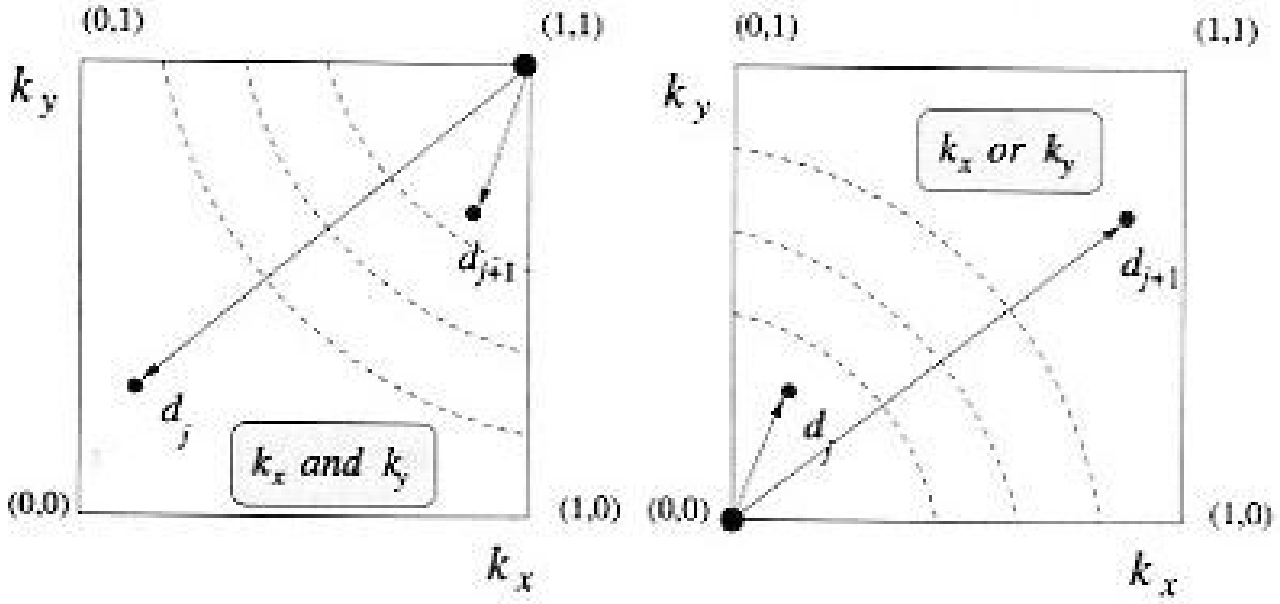
# Fuzzy vyhledávání

- 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




Baeza-Yates, Ribeiro-Neto. Modern information retrieval, s. 39

# 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 | 0.2065 | 0.2368 (+15%) | 0.5573 (+167%) |

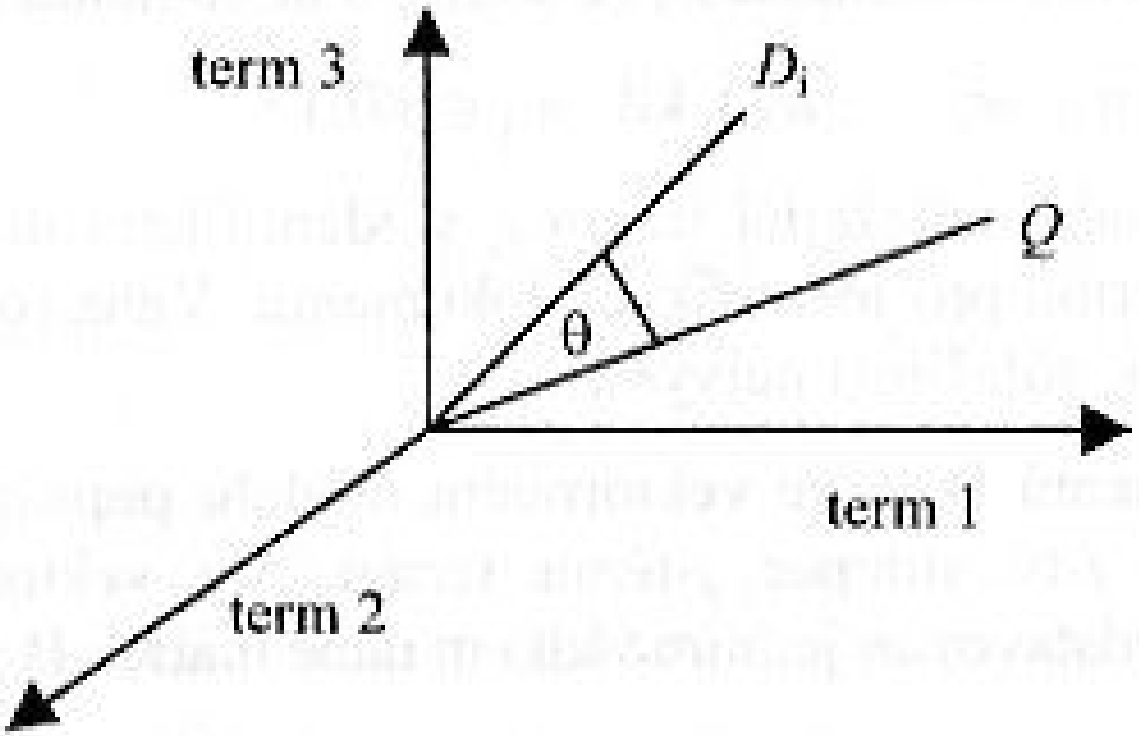
Tabulka 8.5: Srovnání booleovského modelu a jeho rozšíření

Rauch. Metody zpracování informací II. Ukládání a vyhledávání, s. 50

# 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



# 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)