Rešeršní činnost

Mgr. Petr Šmejkal
43262@mail.muni.cz

Information Retrieval

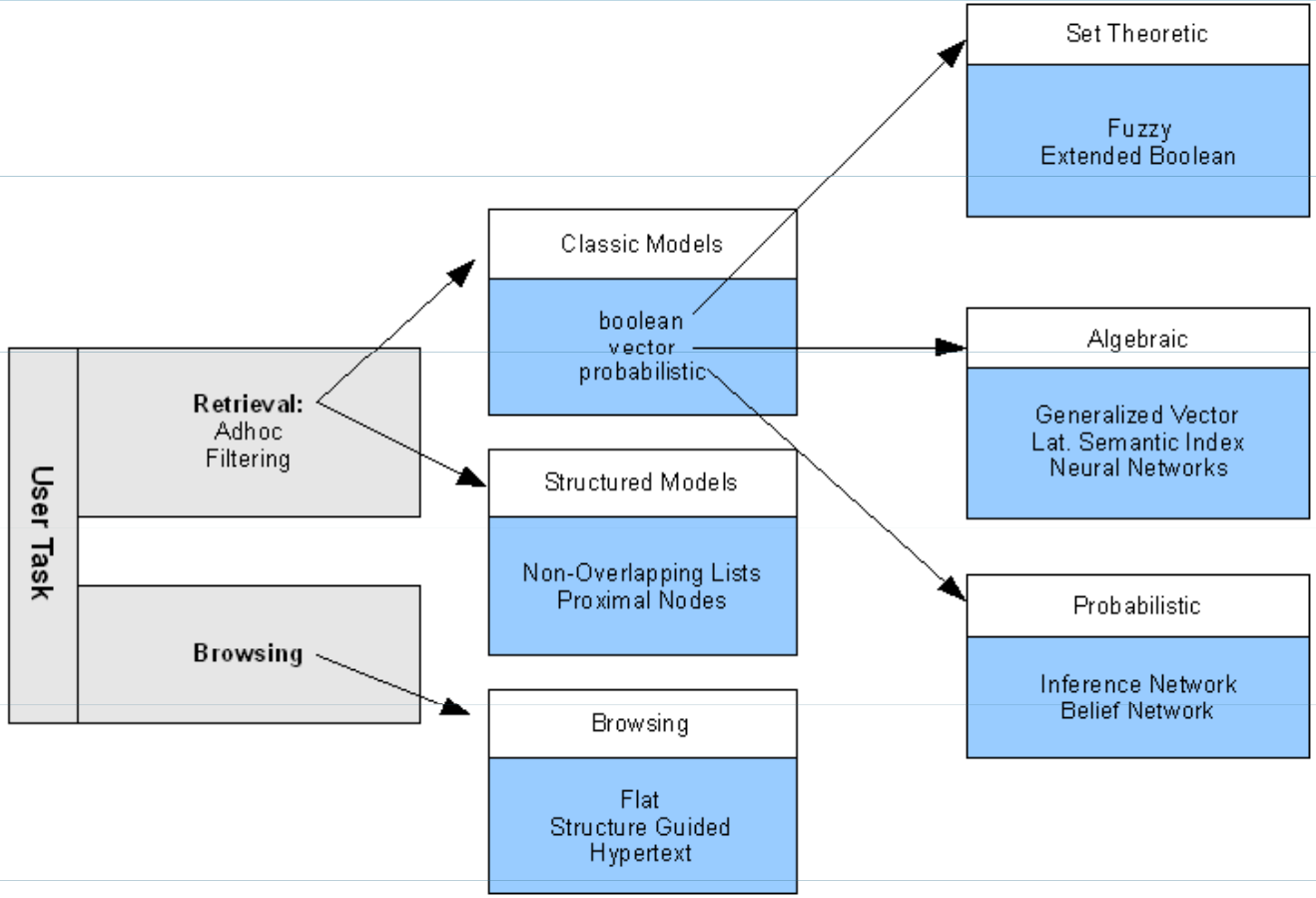
- Vyhledávání informací je činnost, jejímž cílem je identifikace relevantních dokumentů nebo informací v informačních zdrojích (např. fulltextových databázích), souvisí s reprezentací, skladováním, organizací a přístupem k informacím.

- IR je vyhledávání v nestrukturovaných datech.

Vyhodnocování dotazu

- *Precision - Přesnost* : Fraction of retrieved docs that are relevant to user's information need
- *Recall - Úplnost* : Fraction of relevant docs in collection that are retrieved

Mapa IR modelů

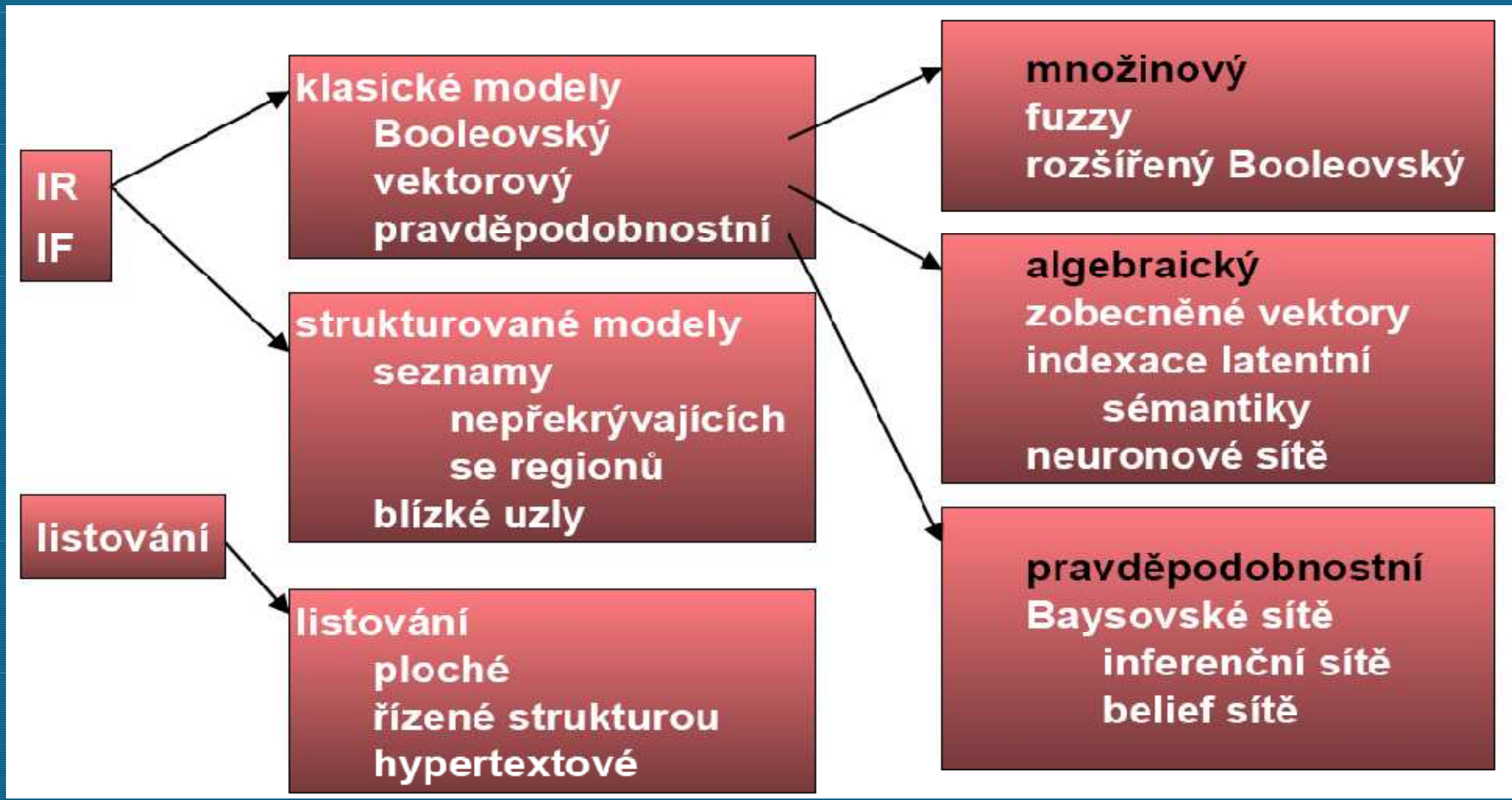


Taxonomy of Information Retrieval Models

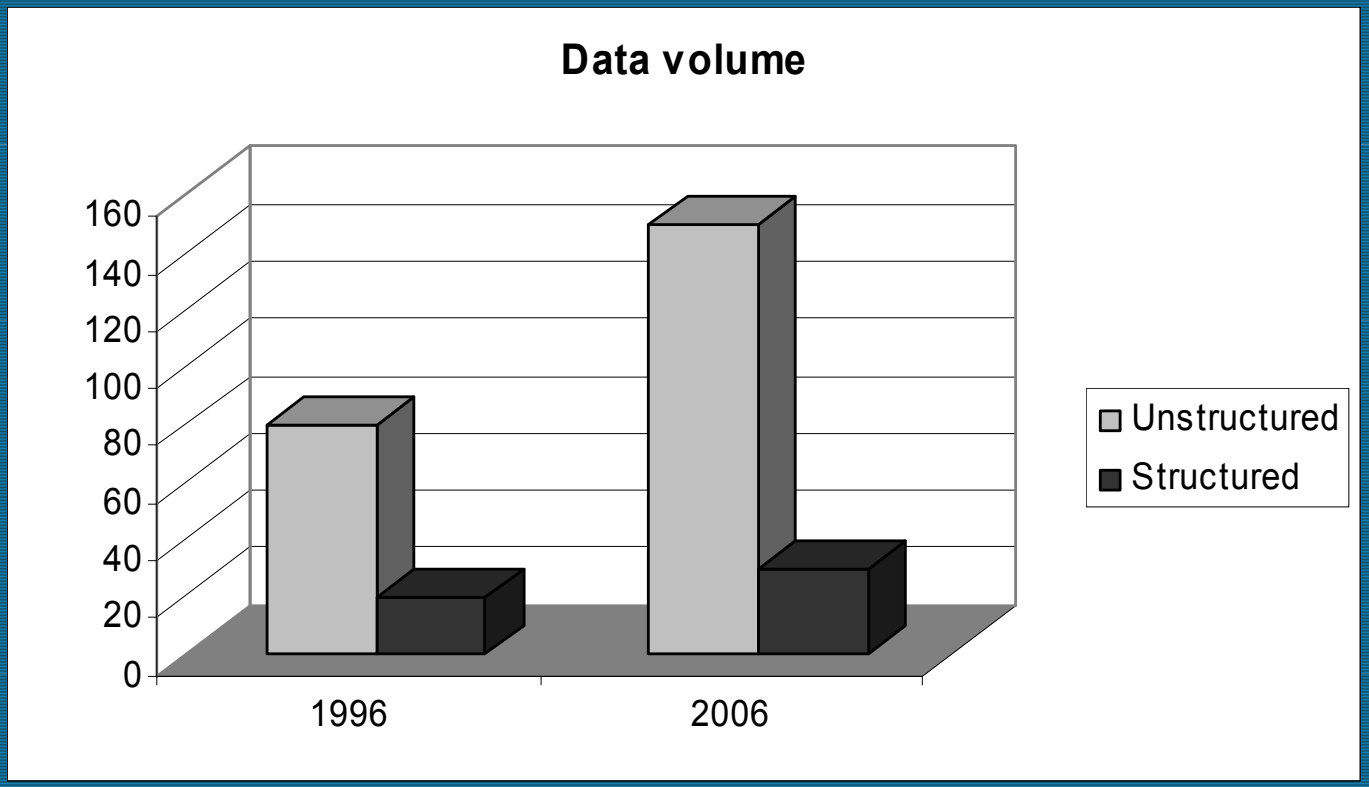
Strukturované modely

- model nepřekrývajících se seznamů (non-overlapping lists)
- sousedních uzlů (proximal nodes)

Mapa IR modelů



Unstructured vs. structured data



- Jiné praktiky a metody
- Vývoj technologií

Unstructured vs. structured data

- Structured data tends to refer to information in “tables”

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

- Typically allows numerical range and exact match(for text) queries, e.g.,
- *Salary < 60000 AND Manager = Smith*.

Nestrukturovaná data

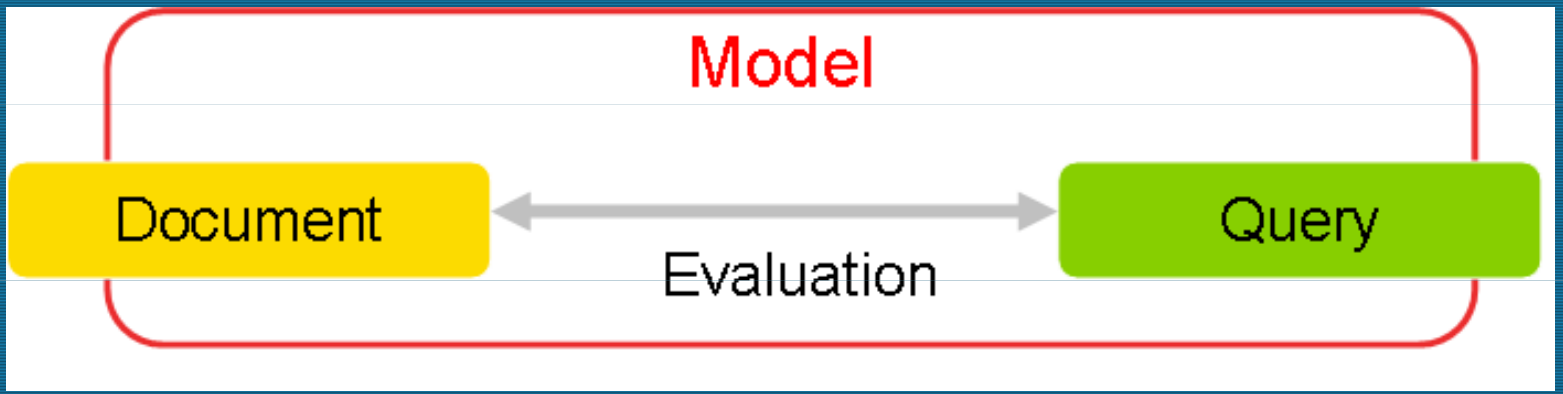
- Většinou volný text
- Lze vyhledávat podle klíčových slov
- Dovoluje i sofistikovanější dotazy – *najdi všechny webové stránky pojednávající o „karburátorech“*
- Pro hledání textu se používají klasické vyhledávací modely

Polostrukturovaná data

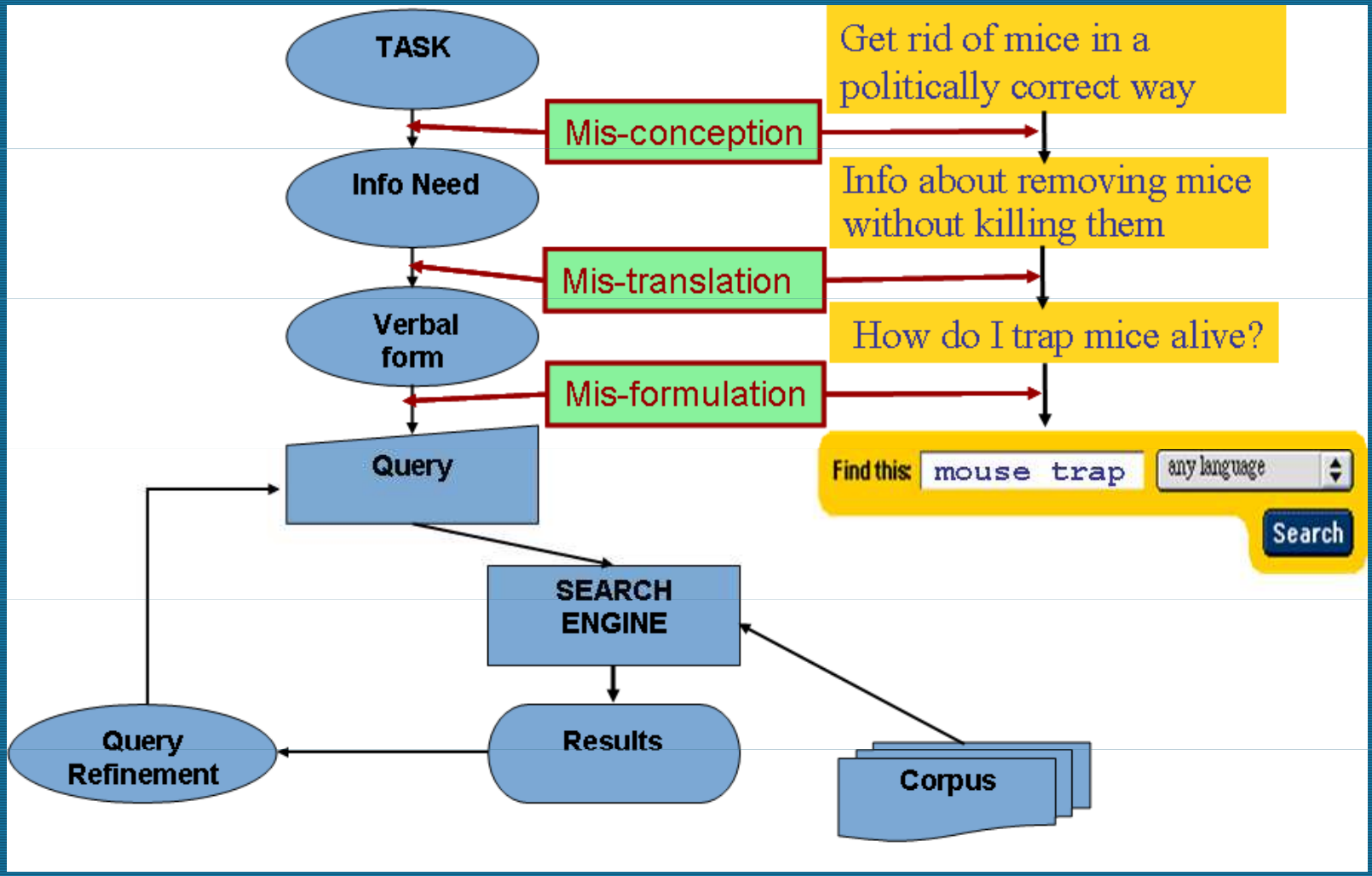
- Většina dat je ve skutečnosti alespoň trochu strukturována (u www např. Title, odrážky, atd…)
- Velké možnosti u XML

IR Modely

- Modely IR odpovídají na otázky relevance dotazu k dokumentům v DB:
  - Jaké dokumenty mají být výsledkem dotazu?
  - Jaké bude jejich uspořádání pro prezentaci uživateli?



Model klasického vyhledávání



Klasický model vyhledávání

Základní koncept

- Každý dokument lze popsat setem klíčových slov nazývaných „index terms“
- „index term“ – slovo, které sémanticky pomáhá lépe určit a pamatovat si hlavní témata dokumentu
- Ne všechny termíny v dokumentu jsou stejně užitečné při popisování obsahu
- Pro lepší orientaci v důležitosti termínů přiřazujeme „váhy“ – vzájemně nezávislé

Booleovský model

- Příklad: která hra od Shakespeara obsahuje slova Brutus a Caesar, ale neobsahuje slovo Calpurnia
- Zápis: ***Brutus*** *AND* ***Caesar*** AND *NOT* ***Calpurnia***

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

- 0/1 vector for each term -> termín buď je 100% relevantní nebo není vůbec – binární systém
- Dotaz buď konjunkce nebo disjunkce vektorů

Charakteristiky modelu

- Komplikovaný překlad – nutná přesná sémantika
- Přes složitost je to nejvíc komerčně využívaný model

Hodnocení modelu

- Výhody
  - Čistě formální systém (vše přesně specifikováno)
  - Jednoduchost
  - Logická jednoznačnost – vždy přesně víme, co bude mezi výsledky
- Nevýhody
  - Přesné zadání dotazu může vést k moc málo nebo k velkému počtu dokumentů

Booleovský model: problémy

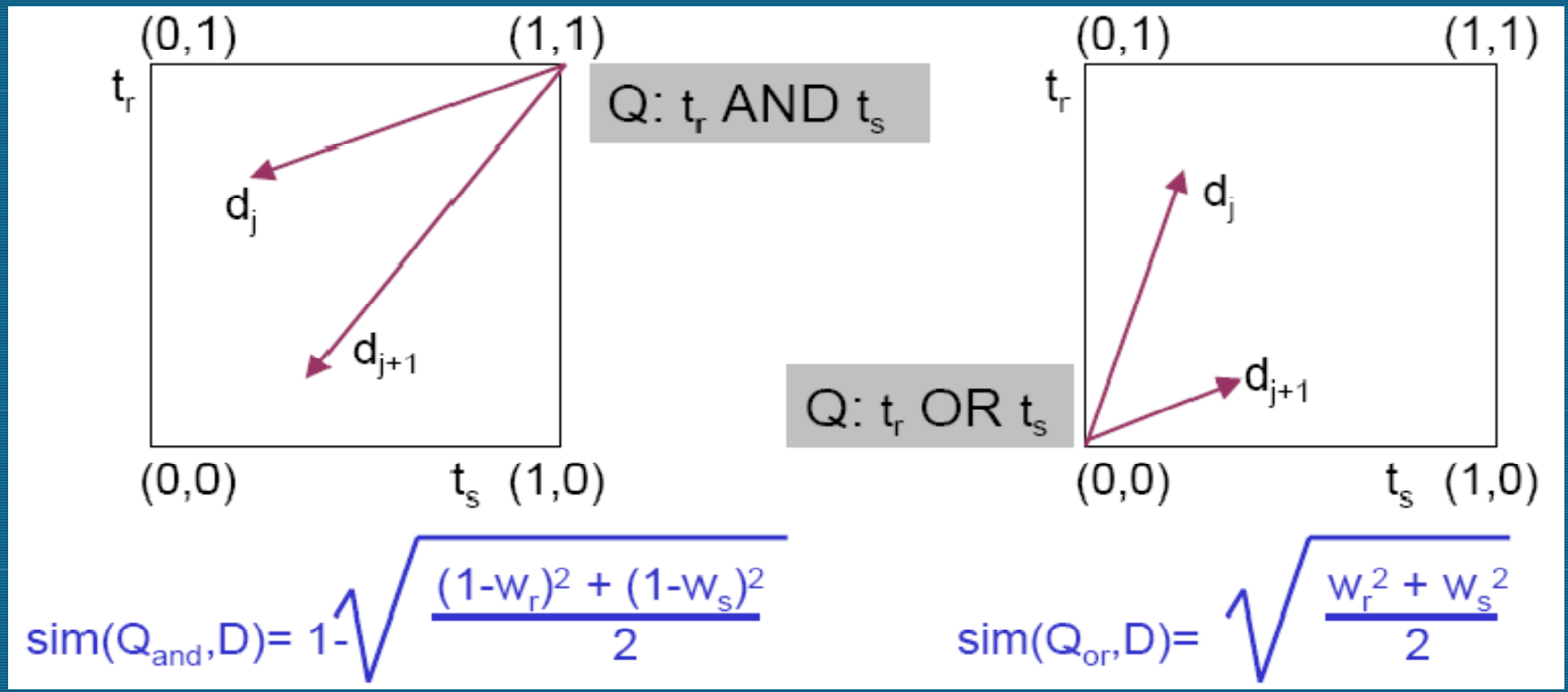
- Kritérium predikce-jak zajistit shodu mezi výběrem termů pro dotaz a dokumenty (dnes: podobnost ontologií)
  - metoda: odstraňováníneurčitosti
- Kritérium maxima – lze zvládnout 20-50 hitů
- Problémy s db úplných textů:
  - velikost db(vs. kritérium maxima)
  - výběr termůpro dotaz
    - přecenění eliminace indexátorů
    - zůstává neurčitost tazatele
  - jednostranné chování tazatele-
    - Tendence měnit poslední rozhodnutí, zachovávat první kroky

Booleovský model: další problémy

- Neintuitivní výsledky
  - A AND B AND C AND D AND E
    - D neobsahujícípouze jeden z uvedených termůnebude vybrán
  - A OR B OR C OR D OR E
    - D obsahujícípouze jeden z uvedených termůjsou chápány jako stejně významné jako dokumenty obsahující všechny uvedené termy.
- Neumožňuje řízení velikosti výstupu.
- Všechny D vyhovujícídotazu jsou chápány jako stejně důležité, není možné je uspořádat podle hodnoty relevance.

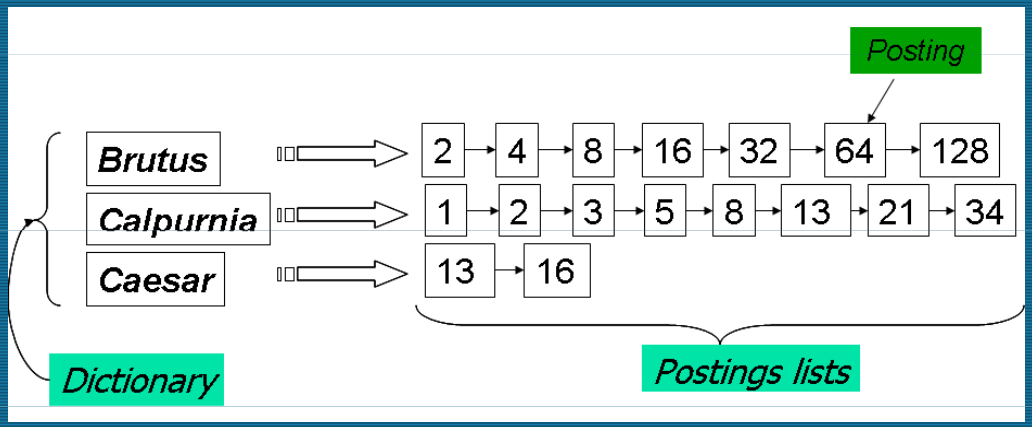
Rozšířená Booleovská logika

- Princip: přiřazení vah termům dokumentů a dotazu.
- Varianta pouze s vahami termů dokumentů $w_t \in <0,1>$.



Invertovaný Soubor/Index

- Pro každý termín zpracujeme seznam dokumentů, které jej obsahují



Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

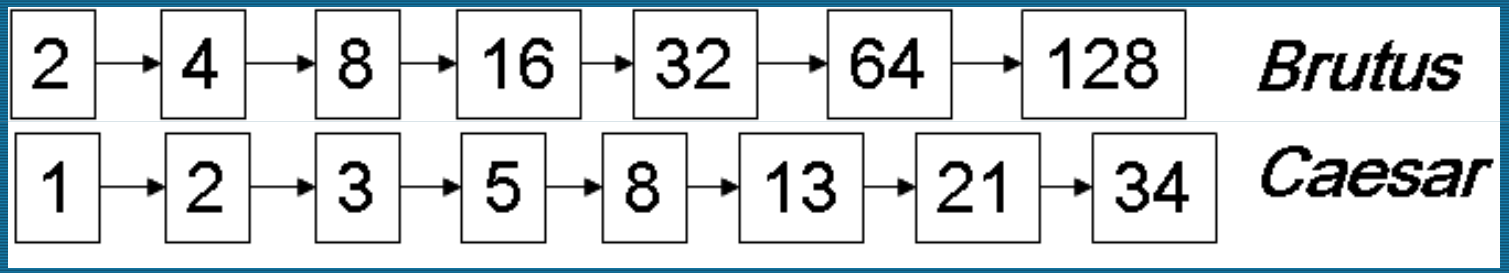
Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Zpracování dotazu

- Příklad: ***Brutus*** *AND* ***Caesar***
  - Locate ***Brutus*** in the Dictionary;
    - Retrieve its postings.
  - Locate *Caesar* in the Dictionary;
    - Retrieve its postings.
  - "Merge" the two postings:



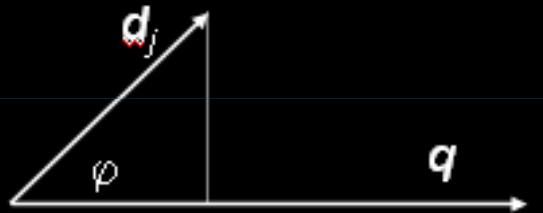
- Výsledek jsou dokumenty 2 a 8

Vektorový model

- Nelineární vektory – může jít až o desetitisíce-dimenzionální prostor
- Počítá se stupeň podobnosti vektorů v dokumentu s dotazem

Váhový vektor přiřazen dotazu q i dokumentům d_j…

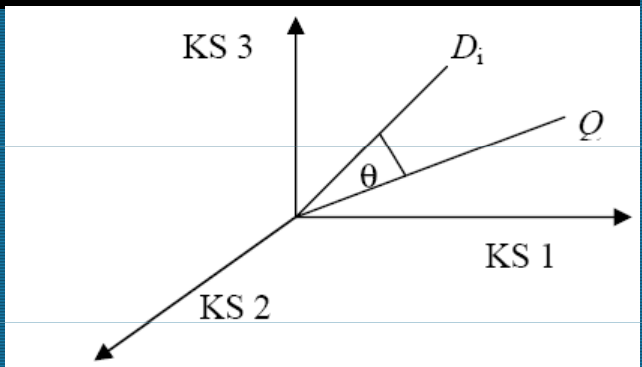
$$\vec{q} = (w_{1,q}, \ldots, w_{t,q}) \qquad \vec{d}_j = (w_{1,j}, \ldots, w_{t,j})$$



pak vzdálenost (kosinová) je

$$\text{dist}(\vec{q}, \vec{d}_j) = \frac{\vec{q} \cdot \vec{d}_j}{\sqrt{|\vec{q}|} \cdot \sqrt{|\vec{d}_j|}} = \frac{\sum_{i=1}^{t} w_{i,q} \cdot w_{i,j}}{\sqrt{\left|\sum_{i=1}^{t} w^2_{i,q}\right|} \cdot \sqrt{\left|\sum_{i=1}^{t} w^2_{i,j}\right|}}$$

ale může být i Eukleidovská, nebo dokonce Manhattan.



Dotaz a dokument v třírozměrném prostoru

Stanovení vektorů

- Předpokládejme, že v textech máme *n* rozdílných slov. Toto *n* také určuje onu n-rozměrnost našeho vektorového systému.
- Každý vektor pak na souřadnici - odpovídající danému slovu - obsahuje jeho četnost (ať již v jednotlivém dokumentu, nebo třeba dotazu).
- Nebývá vhodné nechat růst vektory (jejich délku) nade všechny meze. Proto je výhodné normalizovat používané vektory na jednotkovou délku.

Hodnocení modelu

- Výhody:
  - Zpřesňuje vyhledávací proces
  - Uplatňuje se i částečná shoda s dotazem -> vyšší úplnost
  - Řazení výsledků podle podobnosti s dotazem
- Nevýhody:
  - Předpoklad, že termíny jsou vzájemně nezávislé

Hodnocení modelu

- Jednoduchý model s pružným řazením výsledků
- Systém řazení výsledků je minimálně stejně dobrý nebo lepší než u jiných vyhledávacích modelů
- Přesto, že vektorový model patří k nejstarším, není dosud známá efektivní implementace!
- V literatuře není nikde řešen problém jak postupovat při výpočtu podobnosti dokumentů.

Pravděpodobnostní model

- Předpokládejme, že k dotazu existuje „ideální“ kolekce dokumentů, které tvoří odpověď
- Popisem této kolekce/setu dokumentů můžeme bez problému najít relevantní dokumenty
- Problém je, že vlastnosti ideálního setu neznáme – musíme hádat podle již nalezených
- Dotazování lze chápat jako specifikaci (shlukování) vlastností požadované kolekce dokumentů
- Zlepšování pomocí iterace

Hodnocení modelu

- Výhody:
  - Řazení podle pravděpodobnosti relevance
- Nevýhody:
  - Potřeba „hádat“ první dotaz
  - Nebere do úvahy frekvenci termínů

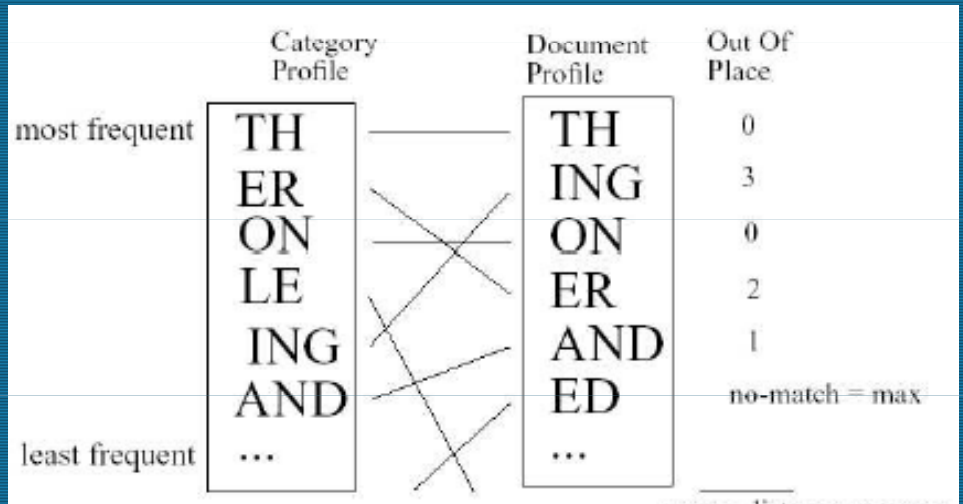
Srovnání s klasickými modely

- Booleovský model neumožňuje uvažovat částečnou shodu – nejslabší
- Salton a Buckley udělali sérii experimentů, ve kterých zjistili že vektorový model dává lepší výsledky než pravděpodobnostní model

Další modely

- Mezi základní modely patří ještě fuzzy model, neuronový, latentní sémantický a spousta modifikací Beyesovských sítí

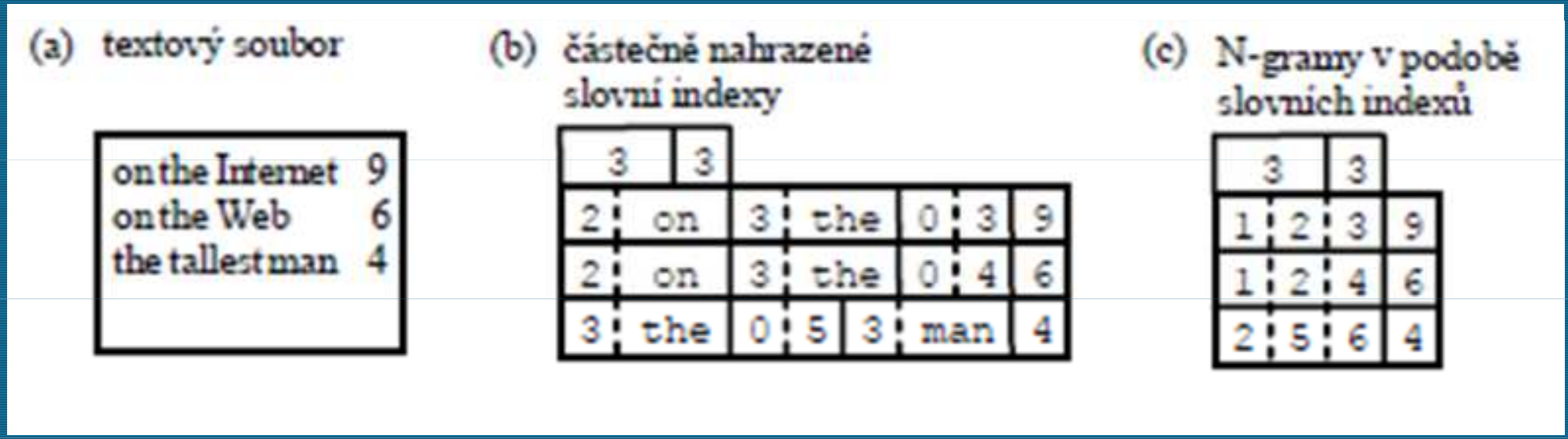
- Jiné než slovní principy:
  - n-gramy
  - lemmatizace



Comparison of n-gram profiles

N-Gramy

- Rozsekání textu na stejně velké části – např. po 3 znacích.
- Vyhledávání podle podobnosti v těchto celcích



N-Gramy

- Vhodné pro zpracování přirozeného jazyka
- Často slouží k rozpoznání jazyka textových dokumentů
- Podobnost s použitím sufixového stromu
- U nás využití např. pro odhalování plagiátů ve vědeckých textech
- Metody postavené na n-gramech jsou odolné vůči posunu textu uvnitř dokumentu