Vyhledávání informací

Operátory, nástroje, strategie

13. 10. 2023

Rozehřívací vyhledávačka

V knize Josefa Lady *Kronika mého života* se prý mluví i o Mahenovi. Rád bych věděl, kde to tam najdu. Můžete mi to najít hned tady od stolu?

Co máme za sebou?

- referenční rozhovor; formulace požadavku
- analýza rešeršního požadavku
- pojmová analýza, tj. hledání klíčových slov, synonym, cizojazyčných ekvivalentů s využitím ŘS atp.
- volba zdroje či zdrojů informací; *query languages*
- volba rešeršní strategie; nástroje, techniky
- formulace rešeršního dotazu

Aplikace operátorů

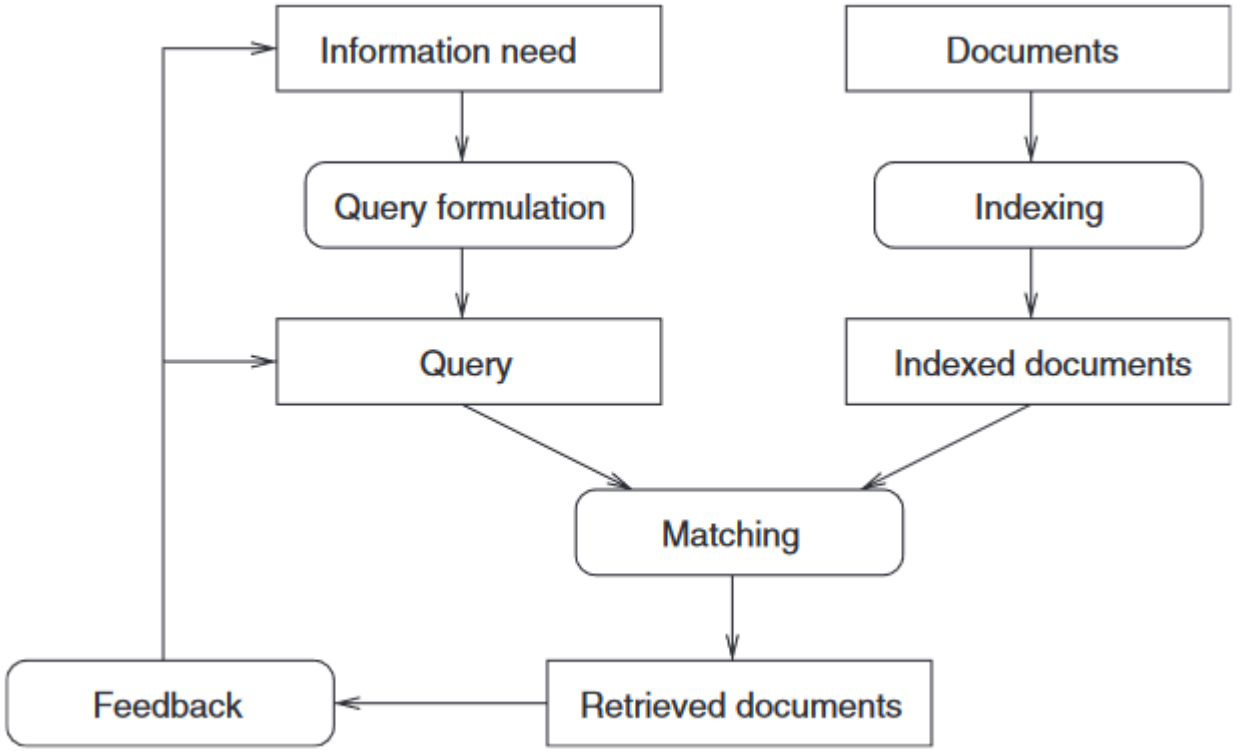
- Boolovské operátory
- *AND, OR, NOT,...*

úkrok stranou

Indexing-based IR

term-based retrieval systems

množina *d*
kolekce *d*
selekční obrazy *d*



Jednoduchý boolovský model

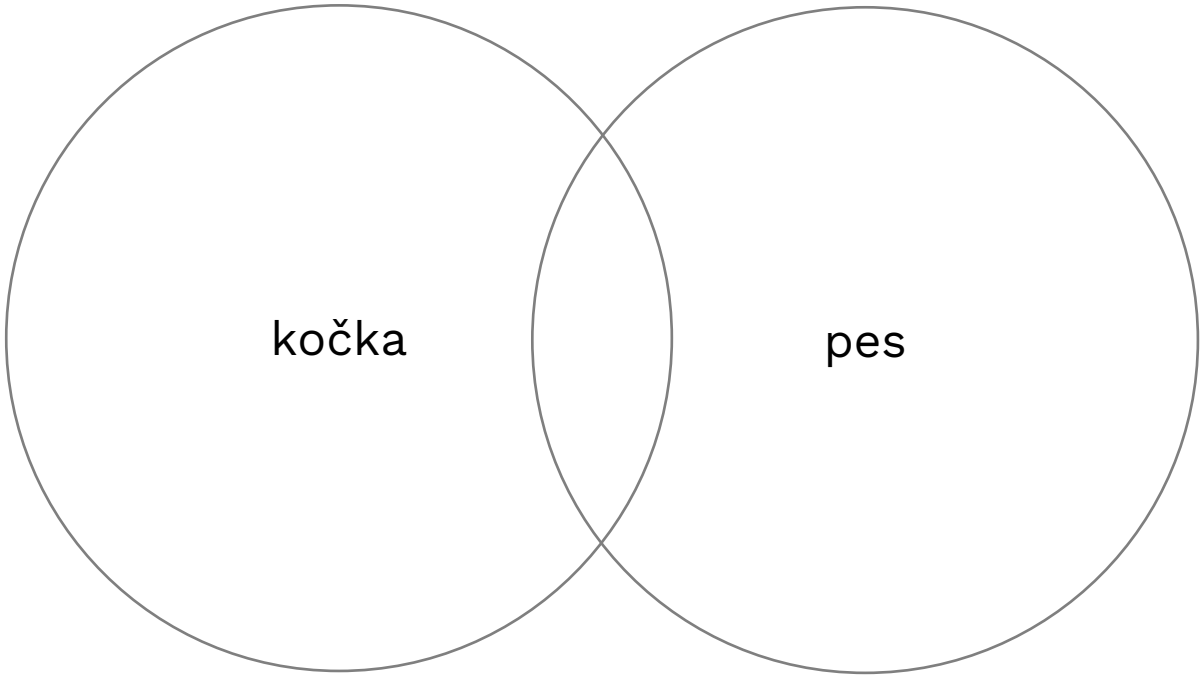
- booleovská logika
- první, nejrozšířenější a široce aplikovaný
- dokument i dotaz jsou pojímány jako soubor výrazů *(term-based)*
- vyhledání je založeno na výskytu výrazů z dotazu

Vyhledávání termínu „jablko“ jednoduše vrací množinu dokumentů, kde je „jablko“.

Pomocí logických operátorů boolovské logiky lze vytvářet nové množiny dokumentů odpovídající vyhledávacímu dotazu.

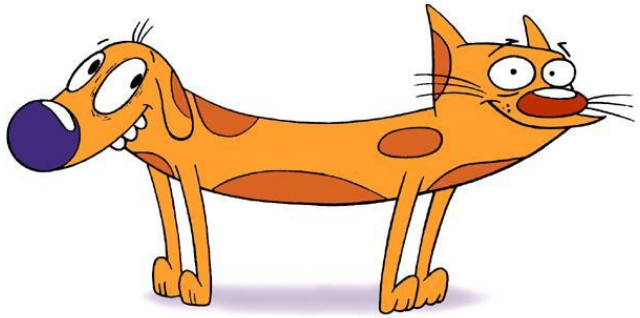
Boolovské operátory

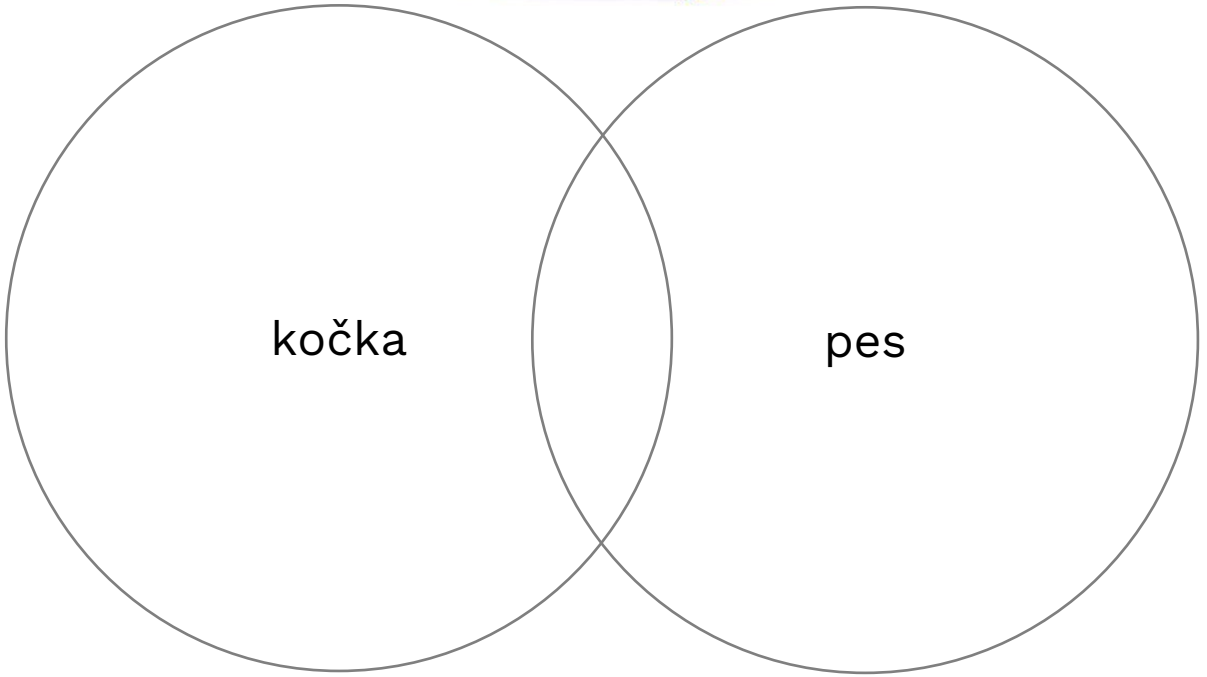
- AND
- OR
- NOT
- XOR

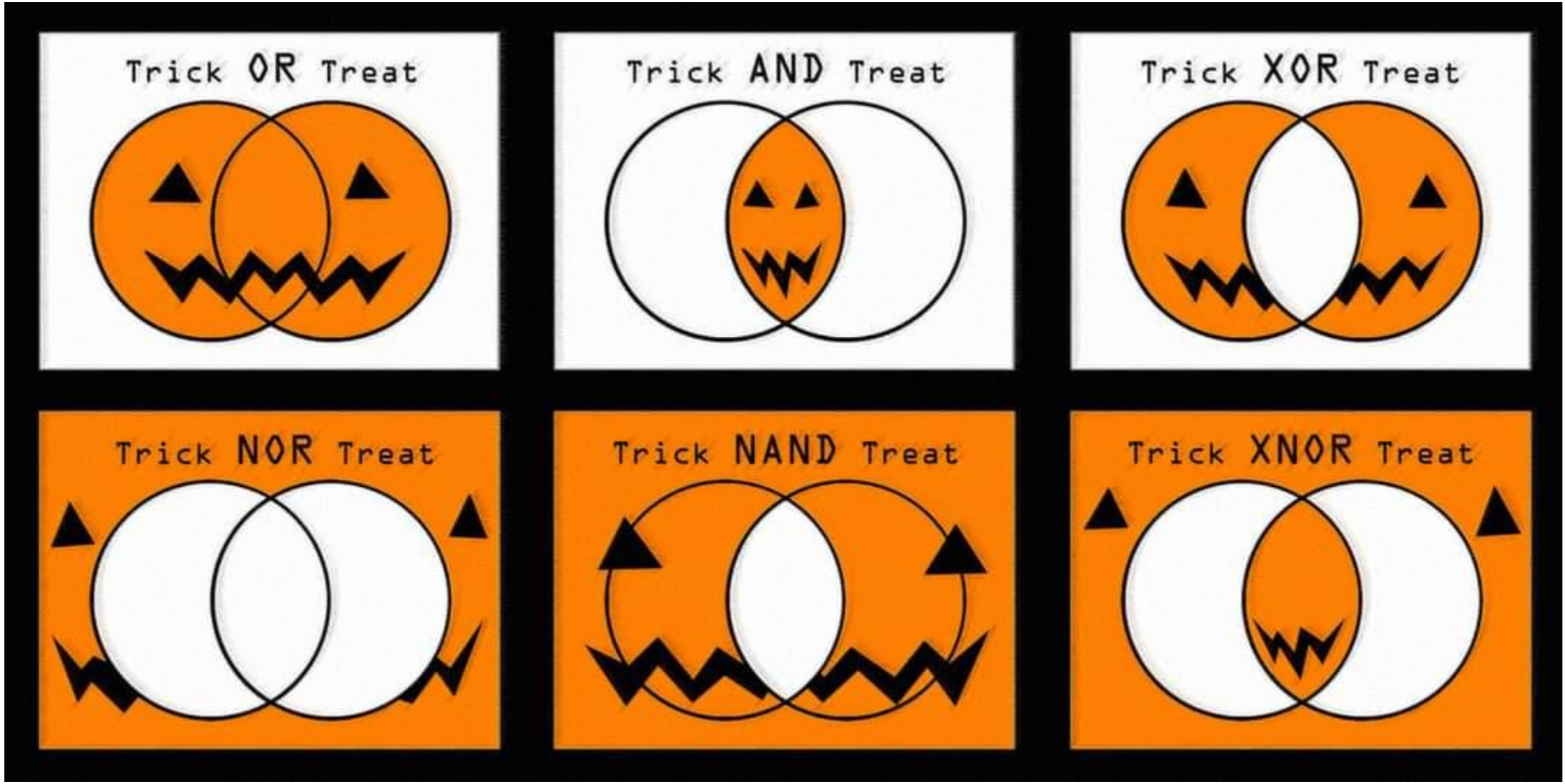


Boolovské operátory

- AND
- OR
- NOT
- XOR






Trick OR Treat
Trick AND Treat
Trick XOR Treat
Trick NOR Treat
Trick NAND Treat
Trick XNOR Treat

((diastolický OR systolický OR krevní) AND tlak) OR ...

malý OR miniaturní OR malinkatý OR drobný OR titěrný

(ipad OR iphone) AND apple

(ipad OR iphone) AND apple

(ipad OR iphone OR macbook) AND apple

Jednoduchý boolovský model

- pocit kontroly vs. značné limity (synonyma, polysémie, kontext)
- výsledky nejsou hodnoceny a řazeny
- model dokument prostě najde nebo ne

„jablko AND sad AND štrůdl“

nenajde „letadlo AND pilot AND létání“,

nenajde „jablko AND štrůdl“ - očividně užitečnější,
ale model nemá jak určit, že tomu tak je…

Modely vyhledávání

- boolovský model
- reprezentace dokumentů skrze vektory
- vektorové modely
- TF-IDF a vážení výrazů
- lematizace, stematizace, stop slova
- indexování latentní sémantiky
- strojové učení, topic modely atp.

Modely vyhledávání

- boolovský model
- reprezentace dokumentů skrze vektory
- vektorové modely
- TF-IDF a vážení výrazů
- lematizace, stematizace, stop slova
- indexování latentní sémantiky
- strojové učení, topic modely atp.

Aplikace operátorů

- Boolovské operátory
- Proximitní operátory
- *různé implementace*
- *smysl jejich využívání na uživatelské úrovni klesá s rozvojem pokročilejších IR modelů a systémů*

Proximitní operátory

- NEAR (např. N3 = kolo favorit, kolo značky favorit, favorit je nejlepší kolo)
- WITHIN (např. W3 = daňová W3 reforma = daňová a sociální reforma, ~~reforma daňová a sociální~~)
- ADJACENT (irská ADJ vlajka = irská vlajka, vlajka irská)
- FOLLOW BY
- PRECEDE BY (např. PRE3)

Vyhledávačka

Používá aleph.muni.cz proximitní operátory?

Pokud ano, jak a jaké?

Vyhledávačka

Používá katalog.muni.cz proximitní operátory? Pokud ano, jak a jaké?

Vyhledávačka

Používá Google proximitní operátory?
Pokud ano, pokuste se ho použít
a vyhodnoťte, jak funguje.

Další vyhledávací techniky

- zástupné znaky filozofie OR filosofie -> filo?ofie
- rozšíření pravostranné, levostranné filo* filo?ofi*
- divoké karty
- vyhledávání celé fráze (jarní vítr -> katalog.muni.cz)
- krácení
- stematizace – nalezení kmene slova *(ProQuest - color)*
- lematizace – vyhledání základního tvaru
- vyhledávání podle polí *(EBSCO -> Field Codes)*
- vážení výrazů

Další vyhledávací techniky

- filtrování
- formální vymezení
- fasetová navigace
- dotaz příkladem
- moderní vyhledávací techniky založené na automatizaci *(JSTOR Text Analyzer)*
- *mnoho dalších…* *(PubChem)*

Každý IZ má svá specifika

Google

- ~~AND~~, OR, XOR, -, ~~AROUND(X)~~
- ~~()~~
- „“
- *
- 0...1
- filtrování
- rozšířené vyhledávání

Každý IZ má svá specifika

- filetype:
- inurl:
- intitle:
- site:
- define:dysfunction
- ~~related:muni.cz~~
- weather:luhačovice
- cache:kisk.cz

google dorking

Každý IZ má svá specifika

Pochopit jak funguje vyhledávání v něm,

seznámit se s možnostmi,

číst nápovědy!

RTFM

Rešeršní strategie

- specifika IZ se odrážejí mimo jiné i ve volbě RS
- rešeršní strategií se rozumí přístup k vyhledávání položek v informačních zdrojích k jednomu požadavku

- strategie stavebních kamenů
- rostoucí perly
- osekávání
- *mnoho dalších strategií*

Stavební kameny

building blocks

Harter (1986)

informační chování seniorů

Stavební kameny

informační chování seniorů

↓

dílčí dotaz 1: informační chování OR informační potřeba

dílčí dotaz 2: senioři OR důchodci OR staří

Stavební kameny

informační chování seniorů

dílčí dotaz 1: informační chování OR informační potřeba

dílčí dotaz 2: senioři OR důchodci OR staří

(informační chování OR informační potřeba)
AND (senioři OR důchodci OR staří)

Stavební kameny

↓

(informační chování OR informační potřeba)
AND (senioři OR důchodci OR staří)

příliš dokumentů? – mohu odebrat pojmy

málo dokumentů? – mohu přidat pojmy

krácení (senio?, potřeb?)

zástupné znaky

Rostoucí perla

snowballing, citation pearl growing

Markey & Cochrane (1981)

- vyhledávání podle nejužšího pojmu z požadavku
- podle nejspecifičtějších termínů / jednoho dokumentu
- cílem je alespoň jeden záznam
- vyhledávání se dále upravuje podle něj (ŘS, terminologie)
- intuitivní postup, který většina z nás dělá

autorský zákon AND pouštění hudby AND restaurace

- autorské právo – produkce hudby

motýl

kopec

množství

hill-topping

Hawkins and Wager (1982) by moji strategii možná nazvali jako „*interactive scanning*„...

A není to tak trochu *berrypicking*?

Osekávání

successive fractions

Meadow & Cochrane (1981)

- dotaz záměrně široký a postupně osekáván
- osekáván pomocí různých taktik
- NOT, filtrace, druh, jazyk, čas

323.1 – etnické menšiny

Rešeršní strategie

- v realitě se tyto metody prolínají, kombinují
- žádná cesta k identifikace rel. zdroje není z podstaty špatná
- v literatuře se kdysi cesty nováčků označovali jako „*naive strategies*“
- dnes se systémy zjednodušují, zpřístupňují
- souvisí to se změnou vztahu *uživatel-informační profesionál*
- mnoho rozhodnutí se automatizuje a schovává do černých skříněk

využiji řízené termíny

využiji stemming

vyhledám jako frázi

zástupné znaky

využiji podřazené termíny v ŘS

vyhledávám podle polí

uvedu synonyma

využiji proximitních operátorů

omezím na určitý typ dokumentu

použiju operátor OR

zkusím NOT pro vyloučení záznamů

řízené termíny použiju jako klíčová slova

upřesním jazykové nebo časové rozmezí

využiji nadřazené řízené termíny

zkusím obecnější termíny s vysokým výskytem

využiji řízené termíny

využiji stemming

vyhledám jako frázi

zástupné znaky

využiji podřazené termíny v ŘS

vyhledávám podle polí

uvedu synonyma

využiji proximitních operátorů

omezím na určitý typ dokumentu

použiju operátor OR

zkusím NOT pro vyloučení záznamů

řízené termíny použiju jako klíčová slova

upřesním jazykové nebo časové rozmezí

využiji nadřazené řízené termíny

zkusím obecnější termíny s vysokým výskytem

Příliš široká formulace

= příliš velké množství výsledků - *nové vymezení tématu?*

- řízené termíny v kombinaci s klíčovými slovy
- podřazené termíny; vyhledávání podle polí
- využití proximitních operátorů
- omezení na určitý typ dokumentu
- NOT pro vyloučení některých záznamů
- jazykové vymezení; časové rozmezí

Příliš úzká formulace

věcné rozšíření

přehodnotit hlediska zúžení

- uvedení synonym; slovních tvarů
- operátor OR, zástupné znaky, stemming
- hledání řízených termínů jako klíčových slov
- nadřazené řízené termíny
- obecné termíny s vysokým výskytem


Mimořádná zpráva

Výuka vyhledávání se 20. října ruší, protože je adapťák.

Zadání závěrečné rešerše

Tabulka ke schvalování témat je v interaktivní osnově.

Téma si vyberte do 10. listopadu!

- 35 záznamů, některé zásadní zdroje mohou být anotované
- primárně ODBORNÉ zdroje - platí zásady podoby rešerše *(příště)*!
- pojmovou analýzu rozepište, popište strategii a postup svého hledání
- využijte alespoň čtyři různé zdroje – uveďte je
- dokumenty musí být alespoň ve dvou jazycích (CZ/SK+X)
- do odevzdávárny nejpozději *4 dny* před termínem zkoušky