

1. DBI007-11 (Paralelní přístup k B-stromům)

2. Hledání dokumentu v kolekci dokumentů

typicky chci najít na disku nějaký soubory, který obsahují (A | B) & C apod.

prohledám všechny dokumenty, předzpracuju je - vyberu z nich významná slova -> trojice (slovo, dokument, pozice v dokumentu)

ten index typicky bude cca 3x větší než původní data - dá se přežít

setřídíme index (vnější třídění)

slovo -> dokumenty... dokument -> pozice...

-> už nepotřebujeme celou kolekci, takže už nejsme větší než kolekce

Samotné hledání

A | B - sloučím stromečky těch dvou slov

A & B - hledám shody

oboje dělám jedním lineárním průchodem, pač ty stromečky těch slov jsou setříděný vzestupně podle všech položek

takhle ale nepoznám, jak moc významný to slovo v tom dokumentu je

problémem je velikost indexu

- vyhodit některá slova - lze jen ve velmi omezené míře

- ta nejčastější (a, že, který...)

- některá co se vyskytují velmi málo

- slova která neexistují apod. - ale inteligentně, musí se brát ohled na text dokumentu apod.

- lematizace - problém u víceznačných slov - statisticky se dá udělat disambiguace; anebo označkovat všechny významy - roste nám zase velikost indexu

- musím znát doménu, nad kterou pracuji

Problémy

zde představená struktura je statická - při změně celé přegenerovat -> dynamizace datové struktury

nemáme odlišeny vícenásobné výskyty, váhy slov - lze přidat

nejen slova, ale i fráze => větší prostorové nároky

Výhody

při hledání nepotřebuju mít samotné dokumenty, na vydání odpovědi mi stačí index

Zde popsaná metoda je pouze základní, dá se na tom stavět (a není zatím známa ta jedna nejlepší metoda).