Posudek diplomové práce

Martin Košalko: „Alternativní vyhledávač systému EGOTHOR"

Cílem práce bylo navrhnout a implementovat alternativní indexační a vyhledávací metodu pro systém EGOTHOR. Autor se zabýval i implementací vlastního systému dotazovaní nad více vyhledávači. Vytvořený systém nabízí otevřené rozhraní pro připojovaní dalších vyhledávačů, respektive metod vyhledávání.

Práci lze rozdělit na dva relativně samostatné celky. První se týká rozšíření systému EGOTHOR o přímou podporu vektorového modelu a tvorbu alternativního vyhledávače FRC, druhý se vztahuje k systému dotazování nad více vyhledávači.

Podpora vektorového modelu

Autor obohatil výchozí index o datovou strukturu, která je transpozicí patřičných invertovaných souborů. To mu umožňuje implementaci potřebných operací pro podporu vektorového modelu.

Na druhou stranu je potřeba uvést, že transpozice probíhá zcela ve vnitřní paměti, což není praktické pro větší indexy velikosti desítek MB a více. Při tomto procesu totiž dochází k tvorbě mnoha malých objektů, které se nemohou vměstnat do paměti běžných PC. Další nevýhodou je, že se v transponovaném indexu ukládají místo číselných identifikátorů jejich textové reprezentace, byť v prefixovém/suffixovém kódování. Zde měl autor poněkud stíženou pozici, neboť číselné identifikace termů nejsou dostupné v původním indexu. Nic mu ale nebránilo tuto číselnou reprezentaci vybudovat.

Transpozici lze provést extrakcí matice termů a dokumentů do n-tic $(term_{id}, doc_{id}, aux)$, kde $term_{id}$ je číselný identifikátor termu, doc_{id} číselný identifikátor dokumentu a aux jsou doprovodné hodnoty (frekvence termu, seznam pozic výskytů daného termu v dokumentu, ad.). V běžném invertovaném indexu jsou tyto n-tice seřazeny nejprve dle prvního a poté dle druhého atributu. Transponovaný index vznikne pouhým přeuspořádáním těchto n-tic tak, aby byly seřazeny nejprve dle druhého a poté dle prvního atributu. Toho lze docílit i pro obrovská data pomocí MergeSort-u v čase $O(n \log n)$. Atribut $term_{id}$ není přímo dostupný (uvažujme třídu `BarrelReader` a rozklad matice do n-tic třídou `Dumper`), EGOTHOR pouze poskytuje invertovaný seznam s textovou reprezentací příslušného termu. Všechny tyto seznamy jsou ale iterovány dle textové reprezentace v A-Z uspořádání, a proto je snadné vytvořit jednoznačné mapování mezi textovou reprezentací termu a číselným identifikátorem.

Je tedy otázkou, zda autor zvolil optimální algoritmus pro samotnou transpozici.

Systém dotazování nad více vyhledávači

V této části se práce zabývá systémem, který by dokázal zvládnout dotazování nad několika vyhledávači s odlišným modelem.

To s sebou přináší problémy v dotazování, protože vyhledávač nad Boolským modelem může nabízet poněkud jiný dotazovací jazyk než vyhledávač nad Vektorovým modelem. Tuto problematiku se ale podařilo v práci úspěšně vyřešit.

Dalším problémem je sloučení parciálních výsledkových listin (PVL). Autor nabídl poměrně jednoduché, ale po praktické stránce vyhovující řešení, kdy je výsledná výsledková listina lineární kombinací PVL.

Vytvořený systém dále nabízí čtyři dotazovací prostředky: logickou formuli, vektor termů, textovou ukázku a nakonec i odkaz na zaindexovaný dokument.

Poslední dotazovací prostředek není de facto implementován, neboť jej nepodporuje ani EGOTHOR ani FRC. Dotazování textovou ukázkou podporuje pouze EGOTHOR, nikoliv FRC. Zbývající dva dotazovací prostředky jsou použitelné pro oba vyhledávače.

Implementace dotazování odkazem by v základním EGOTHORu vedla na implementaci pravděpodobnostního algoritmu pro stanovení množiny podobných dokumentů dle Jaccardova koeficientu, což považuji za úkol překračující rámec této práce.

Domnívám se ale, že dotazování textovou ukázkou v rámci FRC bylo možné realizovat poměrně jednoduše za využití třídy `FTField` s následnou konstrukcí příslušného vektoru metodou `invertize`. I dotazování odkazem bylo snadné, pokud by autor vytvořil datovou strukturu transformující URL na číslo barelu a lokální identifikaci dokumentu s daným URL v tomto barelu. Postačující strukturou by byl například B-strom.

Další připomínky, zpracování práce

Některá tvrzení autora by si zasloužila podrobnější komentář nebo alespoň odkazy na zdroje těchto informací. Na straně 9 je zmíněno, že lemmatizace významně šetří místo v indexu. Toto tvrzení je pravdivé pouze za určitých podmínek (indexujeme pouze slova, lemmatizace slučuje „dostatečné" množství slov, atp.).

Dále se domnívám, že není pravdou, že „... po zpracování h-tého termu..." a „... hodnotách podobností dokumentů větších než $m-h$..." máme „... úplné uspořádání dokumentů ve výsledku..." (strana 31). Podle mého názoru máme pouze dočasné uspořádání dokumentů, které ve výsledku určitě budou.

Bližší vysvětlení by si zasloužilo i tvrzení, že „... barel má takovou velikost, že jej lze najednou umístit do paměti...". Přesnější je, že index je dekomponován na barely tak, aby se odkazy na ně vešly do pole konstatní velikosti (v třídě `Group`). Barel sám o sobě může mít velikost i několik set TB, čímž zcela překračuje byť jen teoretickou kapacitu dnešních vnitřních pamětí.

Dále je otázkou, zda-li je popsaný způsob tvorby profilu (strany 39-41) opravdu takový, že optimálně odpovídá potřebám uživatele (strana 41).

Na straně 27 je uvedeno, že „poslední bit v souboru je nastaven na hodnotu -1".

Práce na některých místech používá nezvyklé obraty, například „termová frekvence termu" (strana 8) místo postačujícího „frekvence termu", „doklad" (strana 39) místo „dokument".

Občas se objevují i překlepy. Strana 20: „soubor termových frekvencí dokumentu (PF)" má být „... (PT)". Strany 23, 44: „... ko(n)verze...". Strana 50: „convaertor_b2e". Strana 52: „DISTTRIBUTOR". Strana 66: „Aut(o)matické".

Vlastní implementace je provedena na odpovídající úrovni. Není ale zcela zřejmé, proč autor vytvářel některé duplicitní třídy, které nabízí výchozí systém EGOTHOR. Jedná se například o třídy pro I/O (package `libs.io` versus `org.egothor.io`) nebo třídu `util.Strings`. Některé z dalších implementovaných tříd mohly být získány děděním z EGOTHORu, například `core.DocMetaData`.

Předložená práce svým rozsahem a zpracováním zcela splňuje nároky kladené na diplomové práce, a proto ji doporučuji k obhajobě.

V Praze, dne 8.5.2006