Posudek vedoucího diplomové práce

Václav Nidrle: Vektorová DIS cartridge pro Oracle

Cílem této práce bylo navrhnout a implementovat cartridge pro Oracle 9, která by umožňovala vektorovou indexaci a vyhledávání v datech.

Autor zvolil řešení v podobě nadstavby Oracle text cartridge s přímým využitím dat, získaných booleovskou indexací. Za přednosti řešení považuji zpětnou kompatibilitu s booleovským vyhledáváním a zahrnutí všech nastavení indexace, které tento model nabízí.

Vlastní vektorové rozšíření nabízí možnost vyhledávání pomocí definovaného váženého vektoru, odkazu na již indexovaný dokument či pomocí volného textu, který je ad-hoc zaindexován a použit jako dotaz. Systém přitom v dobrém smyslu kopíruje rozhraní a postupy z booleovského vyhledávání, takže pro uživatele, zvyklého na originální řešení, je přechod velmi snadný. Řešení nabízí velké množství dodatečných voleb, ovlivňujících indexaci a vyhledávání. Při indexaci je možné zvolit některý z běžných typů výpočtu váhy termu v dokumentu od samotné frekvence termu v dokumentu přes normalizované verze až po zahrnutí charakteristik termů v rámci kolekce dokumentů. Při vyhledávání je možné volit z řady metod výpočtu podobnosti. Jak způsoby vážení termů, tak výpočty podobnosti je možné do systému dále doplňovat.

Práce je psána poměrně kvalitní angličtinou a obsahuje, detailní popis analýzy původního řešení, popis výsledné cartridge z programátorského i uživatelského hlediska a testy výsledného řešení.

Zvolené řešení umožňuje jak vzájemné porovnání implementovaných vyhledávacích algoritmů, tak porovnání s původním řešením, založeným na booleovské logice. To nabízí široké možnosti využití při výuce dokumentografických informačních systémů. Testy, provedené v kapitole 6 ukazují, že:

- Přeformulování booleovského dotazu na volný text vynecháním logických spojek dává v průměru stejné výsledky s booleovským modelem. V některých případech se odpověď zpřesní, někdy je tomu naopak.
- Vhodnou optimalizací vah vektoru bylo vždy možné získat úplnější (obr. 6.4, 6.5, str. 66-67) a zároveň přesnější (obr. 6.6, 6.7, str. 68-69) odpověď.
- Volba podobnostní funkce má nezanedbatelný vliv na kvalitu výsledku. Nejúplnější odpovědi na optimalizované dotazy poskytovala v testech Pseudo-kosinová míra a NTF-ITF metoda vážení termů. V dotazech volným textem bez optimalizace bylo přitom vhodnější vynechat ITF faktor.

Jak vyplývá z testů (obr. 6.1, str. 62), slabinou návrhu je použití standardní vazební tabulky mezi termy a dokumenty pro uložení spočtených vah. Uložení velkého množství řádek do této tabulky a především tvorba standardních indexů nad ní pro pozdější vyhledávání zabírá během indexace velké množství času. Zatímco data vektorového indexu představují zhruba 140% objemu dat indexu booleovského (obr. 6.2, str. 64), vyhledávací B-stromové indexy nad nimi jsou více než dvakrát objemnější. Je otázkou, zda by nebylo výhodnější použít obdobu původního řešení a ukládat váhy binárně po velkých skupinách. I v případě nutného zdvojení uložených informací pro pozdější průchod po řádkách a sloupcích by bylo dat celkově méně a výsledný kód by mohl být pravděpodobně rychlejší. Zdvojení dat by šlo zabránit, pokud by bylo vyhodnocování dotazů implementováno s využitím průchodu po termech.

Celkově se i přes výše uvedené výhrady domnívám se, že práce splňuje všechny kladené požadavky. Doporučuji ji proto uznat jako práci diplomovou.

V Praze dne 11. 9. 2007

RNDr. Michal Kopecký, Ph.D.
KSI MFF UK