Oponentský posudek diplomové práce

Název DP: **Vector IRS cartridge for Oracle**
Diplomant: **Václav Nidrle**

Obsah práce:

Předmětem diplomové práce je návrh a implementace nadstavby (tzv. cartridge) do systému Oracle, pro podporu vektorového modelu DIS. Autor v první kapitole představuje cíle práce, ve druhé popisuje základní modely DIS. Ve třetí kapitole je stručně srovnán boolský a vektorový model. Čtvrtá kapitola obsahuje detailní popis podpory boolského modelu v prostředí Oracle pomocí součásti Oracle Text. V páté kapitole autor představuje vlastní řešení rozšíření systému Oracle o vektorový model DIS, který je postaven na Oracle Text. Šestá kapitola obsahuje experimentální vyhodnocení a sedmá kapitola práci shrnuje.

Hodnocení:

Diplomová práce implementuje plně funkční rozšíření špičkového komerčního produktu, což se na akademické půdě neděje tak často a je třeba to ocenit. Autor provedl zevrubnou analýzu systému Oracle Text a z nabízených řešení implementace vektorového modelu DIS (autor jej nazval Oracle Vector) do systému Oracle zvolil to nejvhodnější z hlediska rozsahu dilpomové práce, tedy využití Oracle Text jako spodní vrstvy. Tím si ušetřil zbytečnou práci na základních rutinách lexikální analýzy, tvorbě slovníků, řetězcových algoritmů, apod. Daní za toto ulehčení je nižší výkonnost a větší prostorová režie, nicméně danému účelu toto nijak nevadí. Celý modul Oracle Vector je implementován jako cartridge – standardní cesta rozšiřování funkcionality Oracle o složitější business datové typy a jejich zpracování. Je implementována celá řada alternativ jak co se týče vážení termů (tokenů), tak i podobnostních vektorových měr. Vážení a měření podobnosti je navíc uživatelsky rozšiřitelné.

Kapitola s experimentálními výsledky obsahuje naprosté minimum, které lze uznat jako dostatečné. Testovací datová sada je velmi malá – 6000 abstraktů o 19000 termech – testování na ní tudíž není příliš reprezentativní. Doporučoval bych spíš kolekce z TREC konferencí, např. Los Angeles Times obsahuje 130000 novinových článků o 250000 termech. Rovněž výběr pouze pěti testovacích dotazů je značně proprietární a nemůže prokazovat typické výsledky. Je obvyklé, že kvantitativní výsledky jsou agregovány jako průměry přes 100-1000 dotazů. V experimentální sekci úplně chybí měření času dotazů. Výše uvedené nedostatky nicméně nepovažuji za vážné, neboť cílem diplomové práce nebyla evaluace vektorového modelu, ale jeho implementace do systému Oracle.

Text práce je psán v angličtině, která je na velmi dobré úrovni. Formálně práce splňuje všechny požadavky. Práce může sloužit i jako referenční/programátorská příručka pro Oracle Text.

Podrobnější připomínky, poznámky:

1) Kapitoly 1, 2, 3 mohly být sloučeny do jediné (1 a 3 jsou krátké).

2) Tabulka DR$INDEXNAME$TF vlastně implementuje invertovaný soubor. Z práce není jasné, jak je na této tabulce definován primární klíč. Pokud by klíčem byla dvojice (TOKEN_ID, Doc_ID), šlo by výrazně zoptimalizovat vyhodnocení dotazu vzhledem k pořadí dokumentů v jednotlivých seznamech (na kterých jsou otevřeny kurzory). Optimalizace by spočívala v předčasném ukončení vyhodnocení dotazu, pokud by počítaná celková podobnost k dotazu klesla pod prahovou hodnotu.

3) Pro vyhodnocení výkonnosti dotazování by nebylo od věci srovnat autorovo řešení s jiným desktop vyhledávacím strojem podporujícím vektorový model.

4) Dotazy definované méně než 10 termy nemají ve vektorovém modelu valný význam. Vektorový model je užitečný především pro „dotazy dokumentem", ve smyslu query by example.

Závěr:

Práce splnila zadání, autor v ní tvůrčím způsobem rozšířil funkcionalitu stávajícího komerčního produktu. Práci doporučuji k obhajobě.

V Praze dne 2. září 2007

Doc. RNDr. Tomáš Skopal, Ph.D.
oponent