

Klady:

- 1) Přehledná textová část.
- 2) Nové řešení v rámci Oracle DBMS.
- 3) Dobrá dokumentace ve zdrojových kódech.


Zápory:

- 1) Vazba na *Précis* bez testů.
- 2) Neúplné porovnání s dalšími systémy (ať už na úrovni rešeršní, nebo experimentální).

Závěr:

Práce splnila zadání. Práci doporučuji k obhajobě.

V Praze dne 17. září 2008



David Hoksza.
Oponent

Oponentský posudek diplomové práce

Název DP: **Index pro textové vyhledávání nad relačními daty**
Diplomant: **Martin Čermák**

Obsah práce:

Předmětem diplomové práce je návrh a implementace fulltextového indexu pro DBMS Oracle s možností indexovat najednou více tabulek, což standardní Oracle Text neumožňuje.

Úvodní kapitoly seznamují uživatele se systémem *Précis*, pro který je možnost fulltextového prohledávání tabulek nezbytná. Další kapitola obsahuje relativně podrobný rozbor dokumentografických informačních systémů a jejich typů. Následuje krátký popis systému pro implementaci vlastních rozšíření Oraclu – *Oracle Data Cartridge* a též popis fungování *Oracle Text* a jeho modifikace pro *Précis* index. Pak již následuje hlavní část práce, která popisuje implementaci vlastního indexu, tj. implementace invertovaného souboru, použité struktury (B-stromy, setříděné seznamy, ...), struktura BLOBů, v kterých jsou uloženy stromy a metadata. V této kapitole jsou též popsány parametry indexu spolu s DB objekty, které jsou metodou využívány. Poslední kapitola je věnována experimentálnímu ověření výkonu implementované metody (vytvoření indexu, jeho aktualitě a vyhledávání v něm) v porovnání se *Oracle Text* a sekvenčním průchodem na dvou kolekcích (CS Wikipedia a FreeDB). Přílohu pak tvoří krátká uživatelská příručka popisující na příkladech práci s indexem v Oraclu.

Hodnocení:

Práce je napsána srozumitelně a je přehledně strukturovaná a prakticky neobsahuje gramatické chyby (až na místy nadbytečné použití čárek).

Vytkl bych nadměrnou vazbu na systém *Précis*, která navozuje dojem, že vyvinutý index bude s touto metodou alespoň testován, když už ne přímo svázán. Nicméně k tomu nikde nedochází a potom nevidím důvod do práce zařazovat sekce o asymetrické váze vztahů mezi tabulkami, váze jednotlivých sloupců atp. Stejně tak názvy prakticky všech struktur a balíčků obsahující prefix *PRECIS* navozuje mylný dojem. Jistě by se dalo nalézt mnoho podobných aplikací potřebujících využívat vícetabulkový fulltextový index a mělo by větší smysl tyto aplikace zmínit.

Dalším nedostatkem, který bych viděl, je neexistence jakéhokoli srovnání s možnostmi v jiných DBMS, ať už s konstatováním, že indexování více tabulek tam není možné, nebo že to možné je a jak (např. MS SQL Server 2005 umožňuje indexovat pohledy, které mohou obsahovat data z více tabulek).

Popis implementace jednotlivých struktur je dostatečně podrobný a kladně hodnotím též zohlednění různých možností implementace jednotlivých částí s ohledem na výslednou rychlost (např. použití více B-stromů pro klíče s různou délkou atp.).

Další kladnou stránkou v aplikaci tohoto typu je důkladná implementace možnosti profilování jednotlivých částí kódu.

V experimentální sekci bylo provedeno porovnání implementovaného indexu s *Oracle Text* i se sekvenčním prohledáváním. Nicméně s *Oracle Text* je srovnán pouze čas potřebný na vytvoření indexu a ne už samotné vyhledávání (i když by bylo omezeno na jednu tabulku).