Posudek vedoucího diplomové práce

Martin Čermák: Index pro textové vyhledávaní nad relačními daty

Cílem této práce bylo navrhnout a implementovat index, který by umožňoval efektivní vyhodnocování dotazů Précis nad danou strukturou tabulek. Základním rozdílem oproti klasickému full-textovému vyhledávání je nutnost prohledávat v každém dotazu obecně všechny sloupce všech tabulek na výskyt požadovaných termů. Ačkoli tak má úloha s full-textovým vyhledáváním mnoho společného, nedá se klasická implementace full-textového vyhledávání, dostupná v nějaké podobě prakticky ve všech databázových serverech použít. Prohledávání celé databáze pomocí operátoru LIKE, použité v DP pana Štullera je pro daný účel velmi neefektivní.

Autor zvolil řešení v podobě cartridge pro Oracle verze 10g (XE i Enterprise Edition). Pro tokenizaci vstupních dokumentů se využívají moduly obsažené v Oracle*Text. Samotná indexace a vyhledávání je implementované samostatně.

Samotný text práce je dle mého názoru přehledně zpracován. Autorovi se v popisu algoritmu podařilo oddělit popis návrhu datových struktur a použitých algoritmů od popisu vlastní implementace těchto algoritmů pro konkrétní databázový server. To usnadňuje případnou reimplementaci pro jinou databázi.

Rozšíření nabízí možnost vyhledávání pomocí booleovských operací. Systém přitom v dobrém smyslu kopíruje rozhraní a postupy ze standardního full-textového vyhledávání. Tuto vysokou míru kompatibility a s tím spojenou uživatelskou přívětivost považuji za jednu z předností navrhovaného řešení.

To, že rychlost dotazování řádově předčí dotazování založené na full-scanem realizovaném vyhodnocování operátoru LIKE nad celou databází, odpovídá očekávání. Zvolené řešení však umožňuje i porovnání implementovaného algoritmu se standardním vyhledáváním nad jedním sloupcem jedné tabulky. Testy měření výkonu, provedené v rámci práce ukazují, že vytváření indexu zabere v porovnání s indexováním stejně rozsáhlých dat pomocí Oracle*Text zhruba dvojnásobek času. Tento handicap by šel pravděpodobně snížit implementací vlastní tokenizace dokumentů. Je však otázkou, zda by v případě rozsáhlých kolekcí dat více nepomohla implementace lepší správy B-stromů při aktualizaci indexu. Především načítání jen těch stránek, které jsou potřeba, a to i za cenu fragmentace dat na více míst. V prostředí Oracle obvykle beztak nelze zaručit, že data budou uložena na disku v jednom souvislém úseku.

Implementace je na CD dostupná jak pro platformu MS Windows, tak pro Linux. Trochu obtížné je se na přiloženém CD vyznat. Uvítal bych v kořenovém adresáři soubor, ze kterého by bylo zřejmé, k čemu disk patří a co se na něm kde najde.

Celkově se domnívám, že práce splňuje všechny kladené požadavky. Doporučuji ji proto uznat jako práci diplomovou.

V Praze dne 17. 9. 2008

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