Oponentský posudek diplomové práce

Název DP: **Modifikace metody Pivot Tables pro perzistentní metrické indexování**
Diplomant: **Bc. Juraj Moško**

Obsah práce:

Předmětem diplomové práce je vytvoření rozšíření metody Pivot Tables (PT), které by bylo optimalizované nejen na počet výpočtu vzdálenostní funkce, jak je běžné u metrických přístupových metod, nýbrž i na počet I/O operací. V rámci práce byla vyvinuta metoda Clustered Pivot Tables (CPT), která je založená na jednoduché myšlence předtřídění dat s využitím M-stromu takovým způsobem, aby byla u sebe v PT data, která jsou i blízko v prostoru. Tuto metodu autor práce nazývá statický CPT index. Dynamický CPT index, pak nepoužívá M-strom pouze pro předtřídění dat, nýbrž ho využívá i dále v průběhu existence CPT jako dodatečnou indexovací strukturu stojící nad PT.

Tímto způsobem je možné významně ušetřit počet přístupů na disk, vezmeme-li v úvahu, že data jsou načítána po stránkách a ne po jednotlivých záznamech. Jako vedlejší efekt vznikla i optimalizovaná verze kNN algoritmu, která umožňuje využít výhody struktury CPT indexu.

Práce je zakončena experimentální sekcí zkoumající počet I/O operací CPT především vzhledem ke klasickému PT indexu. Diskutovány jsou různé možnosti konstrukce CPT a jejich vliv na I/O. Dále je sledován výkon vzhledem k různým datovým sadám a parametrům metody.

Hodnocení:

Asi nejvýraznější výtku mám k rozsahu práce a s tím související kvalitu. Samotný text práce s motivací, přehledem stávajících metod a popisem vlastní metody, zabírá asi 20 stran (bez 3 stránek popisu implementace obsahující popis tříd). Malý rozsah nutně nemusí být ke škodě, ale v tomto případě je to podle mého názoru jednoznačně na úkor čitelnosti. Určitě si dokážu představit rozšířený úvod, kde by motivace měla více než 1 stránku s příkladem užití. Na některých místech jsou uvedeny odkazy na literaturu s popisem některých algoritmů, které se pak používají i v rámci CPT (např. výběr pivotů v metodě LAESA na str. 15). Vzhledem k faktu, že se nejedná o vědecký článek, kde je takový přístup z důvodu omezení velikosti článku běžný, nýbrž o diplomovou práci, měla by být jednotlivým částem algoritmů věnována dostatečná pozornost (místa je zrovna v této práci dost).

Na mnoha místech by čitelnosti práce jistě významně pomohly ilustrace nebo alespoň přehlednější strukturování popisu pomocí seznamů. To souvisí i s prací s formalismy, které by pomohli lepší čitelnosti, nebo struktuře textu. Jediné formální definice v práci jsou vlastně známé definice M-stromu a dotazování nad ním. Nabízí se tedy nad těmito definicemi pracovat při popisu vlastních modifikací spolu s psaným popisem (např. modifikace rozsahového dotazu v sekci 3.5.1).

Práce obsahuje i několik faktických chyb, např.:

- Na konci úvodu je nedokončená věta.

- V definici metrické funkce má být reflexivita definovaná pro jeden objekt O_i a ne pro 2 různé O_i, O_j. Takto je definice metrické funkce chybná.
- U typů dotazů (sekce 1.2) se píše, že rozsahový a kNN dotaz řeší (díky spojení s *query-by-example* modelem) problém deskriptorů, které nejsou samo vysvětlující. To ovšem platí pouze pro kNN, u rozsahového dotazu musí stále uživatel znát sémantiku deskriptorů a vzdálenostní funkci, aby byl schopen vhodně zvolit rozsah.

K experimentální sekci nemám výhrady, spíše naopak – experimenty jsou zpracovány přehledně a všechny metody jsou prozkoumány důkladně s ohledem na jejich smysluplné parametry. Na rozdíl od předchozích sekcí je experimentální sekce zpracována velmi důkladně a kvalitně a jde, myslím, o nejlepší část práce.

Klady:

1) Implementace vlastní indexovací metody.
2) Testování indexu proti stávajícím klasickým indexovacím metodám (PT a M-strom).
3) Důkladná experimentální sekce.

Zápory:

1) Rozsah popisné části práce.
2) Relativně nízká úroveň struktury práce a schopnost formalizace výstupů.

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 19. května 2011

RNDr. David Hoksza, Ph.D.
oponent

Autor/ka: Bc. Juraj Moško

Název práce: Modifikace metody Pivot Tables pro perzistentní metrické indexování

Rok odevzdání: 2011