Oponentský posudek diplomové práce

Název DP: **Externí metrické hašování pomocí D-indexu**
Diplomant: **Jiří Jakl**

Obsah práce:

Předmětem diplomové práce je implementace metrické struktury D-index pro podobnostní vyhledávání ve vysokorozměrných datech a její porovnání s referenčními metodami. Dále byla požadována analýza dynamizovatelnosti struktury a její efektivita pro řetězcové metriky.
V prvních dvou kapitolách autor uvádí do tématu podobnostního vyhledávání pomocí metrických prostorů, představuje některé základní metrické indexační metody, příklady metrik (i nemetrik) a metody volby pivotů. Nejobsáhlejší třetí kapitola popisuje strukturu D-indexu včetně všech možností parametrizace. Analyzuje možnosti dynamizovatelnosti formou automatické volby parametrů nově přidávaných úrovní a popisuje základní operace D-indexu.
Čtvrtá kapitola se krátce věnuje implementaci – popisuje projekt ATOM (implementace referenčních metod) a vlastní návrh implementace D-indexu. V páté kapitole autor testuje experimentálně chování parametrů D-indexu a také provádí srovnávání s referenčními metodami formou kladení stejných dotazů. Poslední kapitola pak stručně shrnuje dosažené výsledky.

Hodnocení:

Celá práce je napsána pečlivě a popisuje celou škálu aktuálních problémů metrického vyhledávání. Vlastní přínos je popsán jak na neformální, tak na technické úrovni. Srozumitelnost formálních zápisů je často podpořena obrázky a příklady, které umožňují rychleji pochopit daný algoritmus či tvrzení. Z experimentů je pak zřejmý vztah k referenčním metodám co do efektivity dotazování. Autor podrobně popsal a otestoval velké množství nastavení D-indexu spolu s jejich vlivem na strukturu a efektivitu D-indexu. Byla také analyzována a následně implementována jejich automatická volba potřebná pro možný algoritmus dynamické konstrukce, čímž byl splněn jeden z požadavků zadání.

I přes rozsáhlost práce a množství prováděných experimentů nad opravdu rozsáhlými a vysokodimenzionálními kolekcemi dat se mohl autor trochu více rozepsat o dopadech nově navrhované metody. S dynamizací je spojená otázka zachování kvality indexu. Dále mohly být prezentovány výsledky porovnání konstrukčních nákladů a velikosti indexů, které jsou nezanedbatelnou součástí každé navržené indexační metody.

Příkladem efektivní dynamizace je např. M-strom (a jeho varianty), který se s rostoucím množstvím dat vypořádává tak, že buduje hierarchii odspodu a tím je zachovávána vyváženost stromu, který nedegeneruje do lineárního seznamu. Pomocí heuristiky štěpení navíc přizpůsobuje velikost regionů nižších pater hierarchie hustotě dat. Navržená metoda dynamizace D-indexu odhadne základní parametry z trénovací podmnožiny dat (počty pivotů a parametry rozdělující funkce pro prvních několik úrovní). Dále rozšiřuje index jen pomocí přidávání dalších úrovní při překročení maximálního počtu prvků v množině vyloučených, kam dle ex-

perimentálních výsledků padá s rostoucí hloubkou daleko méně objektů. Je otázkou, do jaké míry by si D-index zachoval své dobré vlastnosti, kdyby do něj byla většina objektů vložena dynamicky (např. online server na který je denně ukládáno velké množství nových obrázků z internetu). S rostoucí hustotou objektů v jednotlivých kapsách prvních úrovní by pravděpodobně mohla klesnout efektivita vyhledávání v porovnání s referenčními metodami.

Kapitola s experimenty obsahuje velké množství grafů s většinou dostatečným vysvětlujícím komentářem. Nicméně nabízí se pár nezodpovězených otázek k naměřeným výsledkům – proč jsou reálné časy u D-indexu o tolik nižší i v případech, kde je množství výpočtů vzdáleností větší než u referenčních metod? Proč je u obrázku 5.11a pro menší hodnoty k vyšší čas než pro větší hodnoty k?

Podrobnější připomínky, poznámky:

1) Dokument obsahuje na některých místech gramatické chyby, např.

strana	úryvek z textu, chyba zvýrazněna barvou
18	„... naležení do oblasti.“
26	„Obsaženy jsou ale všechna centra ...“
46	„... region řadu n je množina ...“
57	„dvě dvě“
...	

2) Obrázek 2.11 (Hausdorffova metrika) chybně zobrazuje vzdálenost polygonu P2 k polygonu P1 – měla by být stejná jako vzdálenost od P1 k P2.

3) U nově popsané metody v kapitole 2.4.6 (výběr levných pivotů) mohly být více rozepsané konkrétní funkce pro výpočet ceny, na jejichž výsledku je celý algoritmus založen.

Závěr:

Práce svým nadprůměrným rozsahem dostatečně pokrývá všechny body zadání, autor v ní tvůrčím způsobem navázal na předchozí práce na Masarykově univerzitě v Brně. Práci doporučuji k obhajobě.

V Praze dne 26. ledna 2009

Mgr. Jakub Lokoč
oponent