Oponentský posudek diplomové práce

Název DP: **Index pro podobnostní vyhledávání ve vysokodimenzionálních prostorech**
Diplomantka: **Bc. Martina Krejčová**

Obsah práce:

Předmětem práce byla implementace metrického indexu M-index a jeho zapojení do systému Oracle jako data cartridge. Autorka v první kapitole uvádí do problematiky vyhledávání v metrických prostorech, přičemž se zaměřuje zejména na filtrační strategie. Ve druhé a třetí kapitole je popsána struktura a algoritmy M-indexu. V kapitole 4 autorka popisuje návrh integrace metrického indexu do systému Oracle pomocí rozhraní ODCIIndex. V páté kapitole jsou definovány požadavky na uživatele (resp. doménového specialistu), potřebné k implementaci specifického podobnostního modelu (funkce feature extraction a vzdálenostní funkce). V šesté kapitole je podrobněji popsána implementace M-indexu v kontextu Oracle data cartridge, a dále samotné databázové objekty (tabulky a typy) nutné ke zprovoznění indexování a podobnostních dotazů v systému Oracle. Sedmá kapitola uvádí přehled některých dalších metrických indexů a indexační framework MESSIF. V osmé kapitole autorka prezentuje výsledky experimentální studie. Za závěrečnou kapitolou se ještě ve formě přílohy nachází uživatelská příručka.

Hodnocení:

Autorka si zvolila velmi náročnou práci implementačního typu, ve které bylo třeba provést analýzu a implementaci na dvou úrovních – samotná datová struktura M-index a jeho implementace do systému Oracle. Už samotná problematika metrických indexů je značně komplexní, protože se zde uvažují velmi obecné datové modely (autorka se omezila na vektory). Hodnocení bych rozdělil na dvě části. Implementační část práce má nadstandardní rozsah, o kterém svědčí nejen počet řádků kódu, ale zejména příprava potřebná k nastudování problematiky metrických indexů a specifikace Oracle.

Druhou částí je samotný text práce obsahující také návrh implementace. Zde se vyskytuje poměrně dost nedostatků. Předně byl nevhodně zvolen klasický matfyzácký postup "udělej si vše sám, ať to stojí co to stojí". Řeč je o implementaci B+-stromu uvnitř BLOBu ukládaném v tabulce mindex_data. V závěru sama autorka připouští, že tato implementace není efektivní (je třeba deserializovat celý B+-strom). Přitom pro M-index, jehož hlavní datová strukturou je vedle stromu shluků právě B+-strom, by bylo ideální použít nativní zabudovaný B+-stromový index pro sloupec tabulky. V BLOBu by tak byl uložen pouze strom shluků, jehož velikost je relativně zanedbatelná a celá implementace by tak byla škálovatelná a téměř profesionální. V současném stavu je celý index implementován velmi neefektivně a v podstatě nelze použít pro obrovské objemy dat (kvůli kterým má vůbec smysl drahý Oracle kupovat). Experimentální část je velmi chabá a nedovoluje odkrýt nedostatky neefektivní implementace. Ačkoliv je prezentován pouze počet výpočtů vzdáleností (neboť se předpokládá výpočetně drahá fukce vzdálenosti), v experimentech jsou použity dvě maličké datové sady dvourozměrných bodů (a Euklidovské vzdálenosti). Troufám si tvrdit, že ve srovnání se sekvenčním (bezindexovým) průchodem by bylo dosaženo řádově rychlejších reálných časů, než neefektivní implementací M-indexu. Chybí experimenty na vysokodimenzionálních datových sadách (v zadání se hovoří o indexu pro vysokodimenzionální data), pokud možno reálných, atd. Celkově experimentální část práce vykazuje tradiční chyby začátečníka (zde

měl zasáhnout vedoucí). Samotný text práce je srozumitelný, někde je až příliš stručný a nejsou precizně definovány některé pojmy (např. poměrně zásadní termín pivot). Použitá technika výběru pivotů (Incremental) také není jediná možná – existuje jich mnoho a lepších. V kapitole 7 se diskutují 20 let staré metody, přičemž jich existuje spousta nových. Nepochopil jsem uvedení právě frameworku MESSIF, když v něm práce nebyla implementována (takových frameworků existuje spousta).

Přes výše zmíněnou kritiku hodnotím práci jako celek pozitivně, nicméně na tomto hodnocení má rozhodující podíl implementační část práce; jako experimentální studie by samotný text neobstál.

Podrobnější připomínky, otázky, poznámky:

1) Nevhodně členěné kapitoly, sloučil bych 1,7 a 2,3 a 4,5,6.
2) Chybí related work k samotné integraci podobnostního vyhledávání do DBMS, např.
   Maria Camila Nardini Barioni, Humberto Luiz Razente, Agma Juci Machado Traina, Caetano Traina Jr.: Seamlessly integrating similarity queries in SQL. Softw., Pract. Exper. 39(4): 355-384 (2009)
3) Není popsána celá řada použitých PL/SQL konstrukcí, např. co je BFILE v BINDING (BFILE, BFILE)?
4) Celkově je experimentální část velmi slabá, za všechny nedostatky jeden: V tabulce 8.2 se uvádí pro normální rozdělení počet vyhodnocení až 5020 při selektivitě 2907 prvků, přičemž velikost celé databáze je 1348 prvků (kap. 8.3.1). Znamená to, že index vrací duplicity a je 4x pomalejší než sekvenční scan a to dokonce v logických jednotkách?

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 23. srpna 2012

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