Posudek vedoucího diplomové práce

Bc. Martina KREJČOVÁ

Index pro podobnostní vyhledávání ve vysokodimenzionálních prostorech

Cílem práce bylo implementovat index pro efektivní vyhledávání vysoko-dimenzionálních dat. Především pak vhodnou indexovou strukturu a s ní související algoritmy. Žádoucí byla implementace do některé rozšířené relační databáze s tím, že by měla dovolit indexovat pokud možno libovolná data, reprezentovatelná svým feature vektorem.

Autorka na základě studia použitelných datových struktur zvolila víceúrovňový M-index s dynamickým počtem úrovní. Realizace je pak provedena v podobě cartridge pro relační databázi Oracle 10g a výše, která podporuje tvorbu vlastních doménových indexů ve všech svých edicích od volně šiřitelné verze XE až po verzi Enterprise. Konkrétní implementace byla překládána a testována v OS Windows, ačkoli kombinace jazyků PL/SQL a C/C++ nabízí možnost portice a překladu i na jiné, databází podporované, operační systémy jako GNU Linux a další.

Metrický index byl vybraný z více nastudovaných datových struktur, jejichž seznam a stručný popis je uveden v práci. Díky využití objektových rozšíření je index použitelný pro vyhledávání libovolných dat, pro které je jejich vlastník schopen implementovat, nebo specifikovat funkci pro extrakci charakteristik do vektoru čísel (feature extraction), a funkci pro porovnání dvou získaných vektorů, zapouzdřených do objektového typu. Data lze extrahovat z řetězcového typu VARCHAR2 a typů BLOB a CLOB. O vykonání odpovídajících metod při indexaci a vyhledávání se postará polymorfismus, implementovaný v databázi.

Uživatel má k dispozici dva operátory kNN(sloupec,vzor) a range(sloupec,vzor) pro vyhledání *k* nejbližších sousedů, respektive dat v blízkém okolí zadaného vzoru. Pro uživatele je potom vyhledávání podobné například s vyhledáváním v textech pomocí full-textového indexu, který je standardní součástí serverové instalace.

Samotný text práce je dle mého názoru přehledně zpracován. Testování výsledné implementace se věnuje osmá kapitola práce. Porovnává efektivitu vyhledávání (počet porovnání vzdálenosti dvou vektorů) pro různá rozložení dat v prostoru, různé počty prvků a různé počty pivotů. Ačkoli implementace nabízí další možnosti optimalizace, které nebyly prioritním cílem práce, je již nyní poměrně dobře použitelná. Bylo by ale zajímavé otestovat i závislost na dimenzi prostoru, do kterého jsou data mapována. Stávající výsledky jsou k dispozici pro dvourozměrná data.

Celkově se domnívám, že práce splňuje požadavky kladené na práce diplomové a doporučuji ji proto k obhajobě.

V Praze dne 27. 8. 2012

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