Algoritmy pro vyhledání nejbližších sousedů

Erik Hasprunár

Posudek oponenta

Cílem předkládané bakalářské práce bylo nastudovat problematiku vyhledávání nejbližších sousedů a implementovat algoritmus KD-tree v rámci knihovny MESSIF. Dále bylo úkolem studenta navrhnout rozšíření algoritmu pro vyhodnocení kNN dotazů nad více nezávislými kolekcemi a experimentálně vyhodnotit úspěšnost výsledných implementací v porovnání s existující implementací v rámci knihovny VLFeat.

Text práce je strukturován logicky, po stručném úvodu následuje kapitola popisující základy podobnostního vyhledávání, třetí kapitola je věnována struktuře KD stromu, následuje popis implementace a vyhodnocení experimentů. Jednotlivé kapitoly jsou však bohužel velmi stručné, popisy algoritmů jsou neúplné a těžko srozumitelné. Zejména ve druhé a třetí kapitole bych očekávala mnohem důkladnější úvod do problematiky a vysvětlení fungování KD stromu včetně ilustrací. Čtvrtá kapitola popisuje implementaci poměrně dobře, pouze sekce 4.2.5 by podle mého názoru patřila spíše do teoretického popisu algoritmu v kapitole 3. V páté kapitole se nachází výsledky testování, které jsou poměrně zajímavé, ale jejich prezentace je opět slabá. Nedozvíme se pořádně, na jakých datech testování probíhalo, experimenty přeskakují mezi porovnáním dvou implementací a vyhodnocováním vlastností KD stromu samotného. Některé grafy ukazují zajímavé skutečnosti, které ale nejsou nijak diskutovány. Např. podle grafu 5.3 potřebuje autorova implementace méně výpočtů pro vyhodnocení daného dotazu než referenční implementace z knihovny VLFeat – jak je to možné, provedl student nějakou optimalizaci vyhledávacího algoritmu? Naopak přesnost aproximovaného vyhledávání je podle grafu 5.5 ve studentově implementaci podstatně nižší než u referenční implementace – proč?

Po formální stránce mám dále k textu lehké výhrady formátovací (několik sirotků, osamocené předložky na konci řádků) a větší výhradu rozsahovou – čistý text bez obrázků má rozsah cca 35000 znaků, přičemž doporučený rozsah pro bakalářské práce je 54000 znaků. Přitom, jak už bylo výše zmíněno, by důkladnější popis algoritmů byl důležitý nejen pro dosažení správného rozsahu, ale zejména pro zlepšení srozumitelnosti práce.

Implementační část práce je podle mého názoru v pořádku, student se musel seznámit s poměrně rozsáhlou knihovnou MESSIF a v rámci ní vytvořil netriviální implementaci, která se zdá být funkční a dobře strukturovaná.

Celkově se domnívám, že předložená práce splňuje zadání. Implementační část hodnotím jako podařenou, ale textová část podle mého názoru vykazuje poměrně velké množství nedostatků. Předloženou práci navrhuji uznat jako bakalářskou, vzhledem k uvedeným nedostatkům navrhuji hodnocení stupněm **C**. Obhajoby práce se bohužel zúčastnit nemohu.

V Brně, dne 21. 1. 2015

RNDr. Petra Budíková, Ph.D.