Oponentský posudek

Práce: Využití klastrovacích technik při monitorování inzerce
Autor: Tomáš Dzetkulič
Oponent: RNDr. Jan Kára, Ph.D.

Práce Tomáše Dzetkuliče se zabývá metodou klastrování (clustering) a využitím této metody pro klasifikování realitní inzerce. Klastrování je proces, při kterém se dané objekty dělí na skupiny tak, že prvky jedné skupiny mají stejné či podobné vlastnosti. Tato metoda se široce používá v praxi, například při zpracování výsledků fulltextového vyhledávání či kategorizaci obrázků. Použití této metody na klasifikaci inzerce je originální, a tak student musel i samostatně řešit některé problémy specifické pro tuto úlohu.

V první části práce autor studuje obecnou metodu klastrování. Popisuje metody předzpracování objektů, přičemž se převážně zaměřuje na metody předzpracování textu jako filtrování neužitečných informací, rozpoznání jazyka a dělení textu na tokeny. Dále se zabývá způsoby, jak vstupní data transformovat na vektory ve vektorovém prostoru a jak definovat podobnost těchto vektorů. Následně popisuje různé metody klastrování používané při různých problémech. Na konci první části autor porovnává odkazy nalezené pro řetězec "tiger" na různých internetových vyhledávačích. Tyto vyhledávače používají klastrování pro identifikaci různých významů daného slova, odhad jejich relevantnosti a nabídnutí nejrelevantnějších odkazů pro každý význam.

V druhé části práce se autor zabývá aplikací klastrování na identifikaci inzerátů popisujících jednu nemovitost. Nejdříve student detailně popisuje metodu, kterou zvolil pro určení podobnosti dvou inzerátů a dále pak algoritmus, který podle určených podobností vytvoří klastry. Následně se autor zabývá diskusí implementačních detailů popsaného algoritmu a nakonec uvádí výsledky několika testů a volbu parametrů algoritmu.

Celkově v práci autor prokázal schopnost vyhledávání a studia relevantní literatury i schopnost samostatné práce na implementaci nalezených metod a řešení problémů specifických pro klastrování inzerce.

Na druhou stranu jsem v práci nalezl poměrně dost nedostatků technického rázu. Mezi vážnější problémy bych zařadil:

- Práce obsahuje překlepy a pravopisné chyby, i když je psána v rodném jazyce autora.

- Proměnné jsou v textu uváděny normálním písmem, takže se občas obtížně rozpoznávají.
- V kapitole "Miery podobnosti a rôznosti objektov" jedna z uvedených metod využívá kosínu úhlu mezi vektory popisujícími objekty. Uvítal bych u této metody i popis, za jakých podmínek je vhodné ji aplikovat obvzláště proto, že právě tuto metodu autor potom bez dalšího zdůvodnění používá ke klastrování inzerce.
- U popisu metod využitých ke klastrování inzerce chybí jasná definice použité metody pro převod textu na vektory. Tato metoda až později vyplývá z textu.
- Časová složitost zpracování inzerátů je $O(cdmn)$ (ne $O(cdm)$, jak uvádí autor), protože jedna operace nad n-prvkovým vektorem trvá $O(n)$ a ne $O(1)$. V praxi pak n bývá poměrně malé číslo, takže počítače umí s vektory 128 bitů pracovat rychle, ale v teoretickém odhadu by n nemělo chybět.
- U obrázků 8 a 9 chybí vysvětlení, co je na které ose, což činí pochopení kapitoly 6.3 poměrně obtížným.
- V části s testy mi chybí vyhodnocení, jak úspěšná byla aplikovaná metoda na vzorová data.

Drobnější technické problémy jsou pak například:

- U vzorce na straně 13 není popsáno, co je proměnná c_j, ani jak je definováno $d(x_{ij}, c_j)$.
- V kapitole "Suffix tree clustering" by bylo vhodné uvést, jak vrcholy stromu reprezentují fráze (i když si to zkušený čtenář může domyslet).
- V kapitole "Klastrovanie za pomoci zaokrúhľovacích algorimov" mi chybí popis problému PLEB, na který se prý původní problém redukuje. Také detailnější rozebrání výhod a nevýhod tohoto algoritmu by bylo vhodné, vzhledem k tomu, že tento algoritmus autor navrhuje využít k vylepšení jeho současného přístupu.
- Popsaný způsob vytváření klastrů z podobnosti vektorů zřejmě závisí na prvním zvoleném vektoru. Nemůže tato volba výrazně ovlivnit výsledky algoritmu?
- Fukce sim se vzorcem na straně 26 definuje. Formulace, že musí splňovat danou podmínku je poněkud zavádějící.
- U obrázku 6 by bylo vhodné uvést, že se jedná o projekci vícerozměrné situace to roviny ρ.

- Proč se v práci diskutuje implementace třídy pro generování náhodných čísel s normálním rozdělením, když se tato třída nepoužije?
- V kapitole 5.2.6 má být "bit processed nastavený na false" místo na 1.
- Bylo by slušné zmínit, proč se fingerprint již zpracovaných inzerátů nezmění při přidání nového tokenu do slovníku.
- Uváděné naměřené časy běhu jsou pro jak velká data? Pro oněch 10000 inzerátů zmiňovaných dříve?
- U některých odkazů do literatury (např. 12 a 13) chybí časopis, ve kterém byly práce publikovány.

Celkově navrhuji ohodnotit práci známkou 2.

V Praze dne 4. září 2007 Jan Kára