Posudek oponenta diplomové práce

Petr Jenček: Využití vizualizačních metod pro vyhledávání dokumentů

Cílem této práce bylo analyzovat a implementovat nástroj pro zobrazení dokumentů vyhledaných na základě dotazu. Autor problém řešil z pohledu zobrazování nikoli samotných dokumentů, ale zobrazování shluků klíčových slov, charakterizujících nalezené dokumenty a z nich případně odvozených shluků dokumentů.

Pro vizualizaci shluků klíčových slov využil tzv. sociomapování, které zobrazuje vzájemnou blízkost jednotlivých shluků ve dvourozměrném prostoru. Ve 3D zobrazení je možné každému centru přiřadit a zobrazit ještě jednu dodatečnou vlastnost pomocí výšky bodu v terénu.

Práce je napsána srozumitelnou angličtinou, nicméně jsem našel následující nejasnosti v definicích:

- V definici *Idf* (sekce 3.2, str. 7) je symbol *d* (množina termů) definována s využitím stejného symbolu ve významu dokumentu. Bylo by vhodné oba významy odlišit.

- Při popisu shlukovacího algoritmu k-means (sekce 3.3.3.1, str. 15) není zřejmé, jak se stanoví požadovaný počet shluků *k*.

- Je otázka, zda počáteční nastavení, ve kterém se vezme *k* vektorů, odpovídajících *k*-tici v odpovědi neobecnějších termů (sekce 3.3.3.1, str. 15), je opravdu ideální. Pro co nejlepší separaci dokumentů by dle mého názoru bylo vhodnější vzít termy, které současně co nejlépe množinu dokumentů rozdělují na části. Měly by se proto vyskytovat ve zhruba polovině dokumentů a navzájem být pokud možno co nejnepodobnější.

- Práce v sobě zahrnuje informace, získané z řady publikací, používajících různá značení. Srozumitelnosti práce by prospělo značení sjednotit. Například v 3.3.4 se symbol *d* používá pro vzdálenost a nikoli pro dokument, pro rozměr matice s podobnostmi center shluků se používá *N* místo dříve používaného symbolu *k*, a podobně.

Celkově práci hodnotím celkem kladně. Použité metody vizualizace shluků pomocí sociomap nejsou v této oblasti obvyklé a přináší uživateli skutečnou grafickou informaci v podobě 3D terénu. Pro praktické použití by však bylo v budoucnu užitečné více integrovat schopnosti prohlížeče sociomap do vyhledávače, a mít tak možnost kliknutím na daný shluk vybrat z odpovědi odpovídající dokumenty, přeformulovat původní dotaz s použitím odpovídajících termů a podobně.

Přiložené DVD obsahuje externí prohlížeč sociomap a prototypové řešení vyhledávače, který buďto pracuje s vlastní databází textů, nebo přeposílá dotazy jiným vyhledávačům (Yahoo!, Google, ...), a následně shlukuje a vizualizuje termy v odpovědích. Přišlo mi, že „lemmatizátor" v aplikaci se chová mírně odlišně od popisu v práci na str. 30, neboť převádí i domnělý plurál *has* na „singulár" *ha*, zatímco v práci se za plurál považují pouze slova se čtyřmi a více znaky. Samotnou aplikaci zřejmě autor nepovažuje za příliš podstatnou, neboť popisu její implementace věnuje pouze třístránkovou kapitolu *Analysis* na str. 32-34 a na DVD jsem k ní nenašel zdrojové texty. Větší část práce je naštěstí věnována komentovaným ukázkám výsledků pro vybrané dotazy nad různými kolekcemi dat a porovnání kvality shluků pro různé metody výpočtu.

Domnívám se, že práce jako celek splnila požadavky, kladené na diplomové práce. Doporučuji ji proto k obhajobě.

V Praze dne 14. 5. 2009

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