Posudek oponenta diplomové práce

Název DP: **Shlukování textových dokumentů a jejich částí**
Diplomant: **Radoslav Zápotocký**

Obsah práce:

Předmětem diplomové práce (DP) byl návrh a implementace systému pro shlukování dokumentů a jejich částí (tj. celých dokumentů, kapitol, odstavců, vět), a vizualizaci. V první kapitole autor uvádí do problematiky fulltextového vyhledávání a shlukování, ve druhé pak analyzuje funkcionalitu implementovaného systému. Třetí kapitola se věnuje architektuře systému, čtvrtá příkladům použití systému a pátá práci uzavírá. Šestou kapitolu tvoří uživatelská příručka implementovaného systému a sedmou programátorská dokumentace.

Hodnocení:

Práce je přepracovanou verzí práce obhajované v letním semestru. Oproti předchozí byl jak text práce, tak samotná aplikace doplněny o některé funkcionality. V textové části byly zejména přidány podkapitoly 2.1.6 (cluster labeling) a 2.3 (document summarization based on affinity graph), přepracovaná byla kapitola 3 (implementace). Celkově byl text práce rozšířen o 13 na 56. Metody Affinity graph a Cluster labeling byly zohledněny i v kapitole 4 (usage example analysis). Díky přidanému materiálu byl text práce rozšířen a některé původně odfláknuté části přepracovány, také bibliografie byla rozšířena o několik citací. Navzdory tomuto zlepšení práce i nadále působí spíše roztříštěným dojmem; jako ad-hoc výběr metod, které lze provádět nad částmi textového dokumentu, než jako samostatná metoda s jasným účelem.

Dílčí změny doznal také program, do kterého byly jednak začleněny funkce spojené s affinity graph a cluster labeling; jednak autor doplnil prezentaci dokumentu o uživatelsky definované modifikace dokumentu (např. odkazy na podobné kapitoly, apod). Přes tyto úpravy působí aplikace (podobně jako text) stále jako nízkoúrovňový analytický nástroj pro velmi úzce zaměřeného uživatele (vědce) zabývajícího se oblastí Information retrieval. Velmi nepřehledné GUI je samostatnou kapitolou. Užitečnou funkcí pro uživatele-laika by mohly být sumarizační funkce, kdy je celý (!) text abstraktován na základě vět/odstavců/kapitol. Je s podivem, že autora nenapadlo jednoduché rozšíření provádět tyto sumarizace lokálně, tedy po kapitolách či odstavcích. Takto je uživatel odsouzen k nepřesné globální sumarizaci.

Podrobnější hodnocení, připomínky a otázky:

- Aplikace se již nesnaží zapisovat kam nemá, což je jistě správně.
- Cluster labeling je zajímavá věc, nicméně zároveň odhaluje jistou nedokonalost/smysl konceptu, jako příklad 5 „odstavcových clusterů“ z Dobrého vojáka Švejka uvedu:

  stál, Víte, dát, jest, neměl
  hop, tři, neměl, dát, den
  dát, stál, materiál, neměl, den
  dát, dělá, neměl, vypil, li
  stál, tři, set, pět, neměl

  Řešením by bylo zařadit do seznamu stop-slov běžná slovesa aj.

Závěr:

Ačkoliv práce stále působí rozmělněným dojmem, byla doplněna o netriviální funkce a proto ji s výhradou doporučuji k obhajobě.

V Praze dne 22. srpna 2011

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