Posudek diplomové práce:

Shluky silně podobných textů

Autor: Jiří **Diviš**
Vedoucí diplomové práce: RNDr. Martin Holub Ph.D.
Studijní program: Informatika
Studijní obor: Počítačová lingvistika
Praha 2006, 83 str. textu včetně příloh. Součástí je CD s programy + dokumentace

Aktuálnost tématu

Práce se dotýká v dnešní době vysoce aktuálního tématu – z oblasti vyhledávání informací v plných textech, přesněji automatizovanému hledání shluků tématicky podobných textových dokumentů v rozsáhlých textových kolekcích. Aktuálnost tohoto tématu vystupuje do popředí zejména v souvislosti se vznikem nezměrného množství nestrukturovaných informací v podnikové sféře a zejména na internetu. Spojením sítí Intranet na Internet vzniká spojení, které znamená přístup k obrovskému množství dat, ve kterých bude nutno efektivně vyhledávat požadované informace. I když existuje iniciativa *W3C Semantic Web Coordination Group* a množství výzkumných projektů pro usnadnění přístupu k těmto informacím s názvem Sémantický Web, je výsledky těchto snah zasažena jen malá část nově vznikajících dokumentů a to ještě spíše v uzavřených komunitách věnujících se speciální tématice (i když první vlašťovkou je, že v současnosti začínají používat RDF zejména některé zpravodajské servery). Stále je tak prakticky jedinou možností vyhledávání v plných textech, protože neexistuje a zřejmě nebude existovat síla, která by zajistila jednotné zpracování alespoň nových dokumentů umožňující strukturovaný přístup k údajům. Z těchto důvodů je nezbytné se touto problematikou zabývat i v diplomových pracích.

Metoda zpracování

Jednou z metod, která má usnadnit a zefektivnit přístup k relevantní informaci je shlukování. Shlukování jako automatický nástroj pro anotaci dokumentů, pro tvorbu konceptů nebo i jejich efektivní fyzické ukládání. Student v práci použil poznatky z oblasti lineární algebry, matematické lingvistiky, statistiky (shlukování) a teorie a praxe programování. V práci navržené postupy a algoritmy mají být ověřeny experimentálně tedy jde o aplikaci experimentální informatiky.

Přínos práce
Přínos práce vidím v návrhu algoritmu pro shlukování dokumentů a metody pro optimalizaci jeho parametrů na základě učení z dat s použitím aparátu lineárního programování. Zejména oceňuji, že si autor všiml závislosti parametrů na velikosti shluků a porovnal extrapolované a vypočtené hodnoty pro různě velké shluky a porovnal jejich přesnost. Diplomant implementoval a experimentálně ověřil funkčnost navrženého algoritmu. Pro zhodnocení funkčnosti algoritmu použil označkovanou část kolekce českých novinových článků k naučení optimálních hodnot parametrů účelové fce λ, kterou pak použil pro analýzu celé již neoznačkované kolekce. Provedené experimenty naznačují funkční správnost algoritmu i když k jeho ověření by bylo třeba více experimentů, které by umožnily analyzovat jeho základní vlastnosti.

Schopnost vyjadřování a kvalita zpracování

Schopnost vyjadřování diplomanta je poměrně dobrá. Avšak bylo by třeba důsledně zavádět význam symbolů, a sjednotit matematický popis . Bohužel překlepy se dostaly hned už do úvodní části práce. Také odkazy [2,3] na straně 9, které mají odkazovat na monografie o shlukové analýze se minuly cílem. Citace [3].
Dokumentografické systémy (Pokorný a další) jistě nejsou monografií o shlukové analýze. Druhou knihu neznám, ale název tomu také nenaznačuje, že nebude specializována na toto téma.

Připomínky ke způsobu zpracování
V úvodu mi chybí detailnější zpracování současného stavu v oblasti teorie a praxe shlukování, zejména problematika určení optimálního počtu shluků a posouzení kvality shlukování.