Posudek vedoucího diplomové práce

Radoslav ZÁPOTOCKÝ

Shlukování textových dokumentů a jejich částí

Cílem této práce bylo navrhnout a implementovat systém, který by dovolil – na rozdíl od běžných indexačních systémů – analyzovat nikoli kolekci samostatných dokumentů, ale jeden dokument po částech s různou granularitou (například na úrovni kapitol, odstavců, případně vět), vytvářet vektory pro tyto části a ty následně zpracovat, shlukovat, případně zpracovávat pomocí dalších přístupů a výsledky analyzovat.

Oproti první verzi práce doznala mnoha změn k lepšímu ve všech směrech počínaje textem práce, přes programové zpracování a výsledné testování.

Na CD je kromě standardní instalace pro 32bitové operační systémy Windows k dispozici i 64bitová verze, která díky absenci 2GB paměťového omezení zvládá zpracovávat v přijatelném čase i obsáhlé dokumenty s desetitisíci položek (kapitol/odstavců). Při stress testech, prováděných na počítači s 4GB paměti zvládala testovací dokumenty s ~ 20 000 kapitolami. S větší pamětí byla mez posunuta k ještě vyšším hodnotám při době zpracování matice podobností stále řádově v minutách. Díky pokročilejšímu způsobu kešování již získaných mezivýsledků není nutné již spočtené hodnoty počítat znovu a odezva je tak o poznání rychlejší.

Program byl m.j. rozšířen o další alternativy značkování shluků, alternativní výpočty podobností vektorů, a rovněž o třetí – triviální – shlukování. Díky tomu je možné porovnávat výsledky algoritmů, vycházejících ze struktury textu s těmi, standardně používanými pro kolekce dokumentů, značkovat kapitoly pomocí charakteristických slov a vět a podobně. Dalším přínosem jsou rozšířené možnosti zkracování textu výběrem charakteristických vět nebo odstavců a především nově implementovaná možnost přímo do textu doplňovat metainformace, získané během jeho zpracování, a díky tomu uživateli přehlednou formou zpřístupnit výsledky. Implementováno bylo vkládání odkazů na podobné části textu, informace o shluku, do kterého daná část textu náleží, navigace na pokračování textu v daném shluku, odkazy na dostatečně podobné části textu a podobně.

V aplikaci je implementována sada filtrovacích modulů pro zpracování HTML textu počínaje segmentací textu na slova, věty, odstavce a kapitoly, stemming založený na odtrhávání běžných českých a anglických přípon, a vyčleňování stop-slov. Nově přibyla možnost filtrovat slova, nepatřící do seznamu povolených slov, což při znalosti domény dokumentu dovoluje výrazně snížit dimenzionalitu dat a zvýšit kvalitu výstupu. Přítomna je navíc možnost nahrazovat slova jejich významovými ekvivalenty. Vektorizace potom počítá vektory jednotlivých fragmentů pomocí běžného TF*IDF modelu.

Program je řešen modulárně s tím, že si jednotlivé algoritmy doplňují příslušné položky do ovládacích panelů, dostupných v GUI aplikace. Je tak možné relativně snadno doplnit další typy zpracování.

Text práce nyní obsahuje kvalitnější rešerši stávajících algoritmů, které byly v rámci práce použity, případně modifikovány pro zpracování fragmentů textu. V tomto směru by se však práce mohla jít dále a zahrnout důkladnější rešerši možností zkracování textu.

Výsledné dílo tedy v současné podobě představuje poměrně silný nástroj na analýzu dokumentů po částech, vyhledávání navzájem si podobných částí, jejich shlukování. Výsledky mohou být promítnuty přímo do HTML kódu dokumentu, nebo exportovány v řadě formátů (CSV, PNG, HTML) pro ruční či strojové zpracování v nástrojích třetích stran.

Za přínosné považuji možnosti zkracování textu pomocí modifikace algoritmu pro výpočet Affinity grafu, kdy je možné volit libovolně krátkou reprezentaci s tím, že se postupně vybírají fragmenty textu (věty nebo odstavce), které nejlépe reprezentují největší části obsahu textu.

Na přiloženém CD jsou sice čtyři testovací dokumenty, včetně pracovní verze diplomové práce, ale testy v práci jsou provedeny na jednom dokumentu. Bylo by lepší provést testy na více typech dokumentů s rozdílnou délkou a zaměřením (manuály, vědecké články a publikace, elektronické sborníky) a výsledky pro různé typy porovnat. Závěr by pak mohl být reprezentativnější. Domnívám se, že po provedení takových testů a důkladnějším

porovnání výsledků jak navzájem, tak s výsledky obdobných algoritmů, by bylo možné výsledky práce publikovat na vhodné konferenci.

Celkově se domnívám, že autor prokázal schopnost tvorby rozsáhlejších programových děl, stejně jako schopnost tvůrčím způsobem rozvíjet stávající přístupy a algoritmy pro zpracování textu. Doporučuji proto práci uznat jako práci diplomovou.

V Praze dne 15. 8. 2011

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