Oponentský posudek disertační práce

Program: FI D-IN4 Informatika
Obor: FI IN Informatika
Student: RNDr. Jan Pomikálek
Název práce: Removing Boilerplate and Duplicate Content from Web Corpora

Oponent: RNDr. Pavel Pecina, Ph.D.
Pracoviště: Centre for Next Generation Localisation, Dublin City University, Irsko

Výsledky práce

Předkládaná disertační práce Jana Pomikálka se zabývá velice aktuálními tématy souvisejícími s vytvářením tzv. webových korpusů, tedy korpusů, které tvoří automaticky zpracované texty z webu. I přesto, že sám autor tvrdí (str. 12), že texty z webu jsou „a perfect resource for building text corpora", tato práce je důkazem toho, že využití webových textů pro budování korpusu je do značné míry problematické (nemluvě o právní stránce věci, kterou se však práce nezabývá).

Autor v práci řeší dva snad nejpalčivější problémy, které s budováním webového korpusu souvisejí: čištění webových stránek od nepotřebného obsahu a odstraňování duplicitních a téměř duplicitních textů. Pro obě úlohy platí, že nejsou nové – zejména v posledních letech jim byla věnována značná pozornost, a to jak v praxi, tak literatuře – Jan Pomikálek ale svou disertační prací významně přispívá k jejich řešení: pro čištění webových stránek navrhl jednoduchý, ale při tom účinný, heuristický algoritmus jusText, pro detekci a odstraňování duplicitného textu vylepšil a především výrazně zefektivnil metodu založenou na analýze výskytu duplicitních n-gramů.

Autor provedl, popsal a analyzoval celou řadu experimentů, což je jedna z nesilnějších stránek práce. V případě čistění webových stránek byly experimenty věnované optimalizaci parametrů navržené metody a její důkladné (intrinsické) evaluaci a srovnání s ostatními přístupy. V případě detekce a odstraňování duplicitního textu byla většina experimentů zaměřena na analýzu chování algoritmu na různých korpusech. Další ze silných stránek práce jsou rozsáhlé a důkladné rešerše a analýzy existujících metod, které mohou spolehlivě sloužit jako výchozí bod pro případné další zájemce o tuto problematiku.

Výsledky práce byly buhužel publikovány pouze formou jednoho (ne plně recenzovaného) příspěvku na (spíše podprůměrné) konferenci LREC a jistě by si zasloužily větší publicitu.

Struktura práce

Práce je vhodně strukturovaná, členěná do čtyř kapitol a dvou příloh. První kapitolu tvoří úvod do problematiky webových korpusů a motivace celé práce. Jádro je v kapitole druhé, která pokrývá problematiku čistění dat, a v kapitole třetí, ve které se autor zabývá detekcí a odstraňováním duplicitního obsahu. Výsledky práce jsou shrnuty v kapitole čtvrté. Přílohu A tvoří snímky aplikace korpusového manažeru zobrazující výsledky vyhledávání ve webovém korpusu před a po aplikaci metod prezentovaných v disertační práci. Příloha B prezentuje výsledky evaluace těchto metod formou přehledných grafů.

Jazyková a grafická úroveň

Práce je psaná anglicky, obsahuje jen minimum pravopisných chyb a překlepů. Text je dobře čitelný a srozumitelný. Grafická úroveň práce je také v pořádku.

Připomínky k textu

Seznam použité literatury je relativně obsáhlý (obsahuje přes 60 publikací), ale reference v textu často chybí. Nezřídka se v práci vyskytují neopodstatněná, často i odvážná a těžko prokazatelná tvrzení, u kterých není jasné, zda jsou převzatá (případně odkud), nebo se jedná o vlastní autorova pozorování (případně na základě čeho), např: „Though they [part-of-speech taggers] are not hundred percent accurate they are adequate for most aplications." (str. 12), „Comparable corpora are much easier to create then parallel corpora" (str. 14), „ ... most of the of the duplicate and near duplicate pages occur within the same [web] domain"

(str. 75). Nevhodné jsou reference na stránky Wikipedie, zvlášť v případech, kdy je možné se odkázat na renomovanější zdroje (např. „Vector Space Model", str. 56). Reference místy chybí také v případech, kdy by bylo žádoucí použitý termín vysvětlit přímo v textu, např. „Hamming distance" (str. 59).

Podobně v části 2.7, kde autor popisuje algoritmus jusText, jsou uvedeba čtyři pozorování, která jsou zásadní pro návrh tohoto algoritmu, ale vůbec není zřejmé, z čeho vycházejí, na jakých datech byla provedena a do jaké míry jsou platná.

Otázky k obhajobě

Pokud bych měl práci něco vytknout a položit dodatečné otázky, týkaly by se následujících bodů:
1) Algoritmus pro hledání duplicitních a téměř duplicitních textů je založen na uchování informace o duplicitních n-gramech a je tím velice vhodný pro nasazení na pevně dané (neměnné) množině dokumentů. Tato vlastnost jej ale znevýhodňuje pro použití pro inkrementální budouvání webového korpusu. Vzledem k množství duplicitního obsahu na webu je spíše žádoucí duplicitní a téměr duplicitní dokumenty identifikovat hned po stažení z webu, bez nutnosti je (všechny) ukládat a zpracovávat až následně. Je možné použití algoritmu i tímto způsobem?
2) Algoritmus jusText navržený pro čištění webových stránek má několik parametrů, které je nutné nějakým způsobem optimalizovat. Postup navržený a prezentovaný v práci je více inženýrský něž vědecký, spočívá v ručním ladění jednotlivých parametrů a vyžaduje ručně anotovaná data. Autorovi se sice podařilo nastavit parametry tak, že výsledky algoritmu na testovacích datech jsou v srovnatelné, v někerých aspektech dokonce lepší než výsledky ostatních metod, ale není bohužel jasné, do jaké míry je nastavení parametrů závislé na použitých datech, co by vyžadovalo použití na datech lišících se v jazyku či doméně, jak velká trénovací data by byla v takových případech potřeba, a jakou úspěšnost by bylo možné ve výsledku očekávat.
3) Značná část práce je věnována intrinsické evaluaci a srovnání úspěšnosti metod pro čištění webových stránek. Text obsahuje množství tabulek a grafů, ale výsledky se v mnoha případech liší jen nevýrazně a vzhledem k tomu, že nebyly provedeny statistické testy signifikance, lze o významnosti některých rozdílů pochybovat.
4) (Intrinsická) evaluace metody pro hledání duplicitních a téměř duplicitních dokumentů ve webovém korpusu v práci chybí. Byla sice provedena rozsáhlá případová studie na části korpusu ClueWeb09, ale čtenář se v podstatě nedozví, jaký efekt má použití algoritmu, co se kvality výsledku týče. Z části je to možná proto, že si autor nestanovil měřítko takové kvality, a to také souvisí s tím, že v práci není blíže definováno, co lze považovat za téměř duplicitní obsah a co již ne (i přesto, ža takové definice bude závislá na konkrétní aplikaci, viz dále).
5) Další nedostatek vidím také v chybějící extrinsická evaluaci. Autor kladl velký důraz na úspěšnost a efektivitu dílčích metod, ale v podstatě žádný prostor není věnován analýze použitelnosti výsledného korpusu v nějaké konkrétné aplikaci ani srovnání s korpusy vytvořenými klasickou cestou.

Závěr

I přes uvedené nedostakty nelze pochybovat o tom, že práce Jana Pomikálka je kvalitní a je významným příspěvkem v oblasti vytváření webových korpusů. Lze také bez obav konstatovat, že:
1) Autor ve své disertační práci prokázal schopnost samostatné tvůrčí práce v daném oboru.
2) Práce splňuje požadavky standardně kladené na disertační práce v daném oboru.

RNDr. Pavel Pecina, Ph.D. 10.8. 2010, Dublin