Oponentský posudek disertační práce

Západočeská univerzita v Plzni, Fakulta aplikovaných věd
Obor: Kybernetika
Studentka: Ing. Lucie Skorkovská
Název práce: Vyhledávání informací v řeči a využití slepé zpětné vazby

Oponent: RNDr. Pavel Pecina, Ph.D.
Pracoviště: Ústav formální a aplikované lingvistiky, MFF UK, Praha

Obsah práce

Předkládaná práce Ing. Lucie Skorkovské se zabývá metodami vyhledávání informací aplikovanými na řečová data, což je oblast nabývající v poslední době stále většího významu. Autorka se zabývá především metodami slepé zpětné vazby, o kterých je známo, že v případě tradičního vyhledávání v textových datech často přinášejí významné zlepšení výsledků, tedy kvality vyhledávání. Autorka se ve své práci pokouší ověřit, jakých výsledků lze dosáhnout aplikováním těchto metod na mluvenou řeč, kterou je nejdřív nutné automaticky přepsat do textové podoby, což je typicky chybový proces, a obohacení vyhledávacích dotazů o nesprávně rozpoznaná slova může vést ke zhoršení kvality vyhledávání. Významnou součástí práce je vylepšení celého postupu automatickou predikcí optimálního počtu dokumentů pro zpětnou vazbu, a to pro každý dotaz zvlášť. Typicky se hodnota tohoto parametru stanoví bez ohledu na dotaz. Předkládaná práce obsahuje popis velkého množství experimentů a jejich evaluace včetně diskuse výsledků.

Struktura práce

Práce je vhodně strukturovaná, členěná do 8 kapitol, opatřená seznamem citované literatury a bibliografií autora. První kapitola obsahuje stručný úvod celé práce a především specifikaci jejích výzkumných cílů. Druhá kapitola obsahuje jednak teoretický úvod do problematiky vyhledávání informací v řeči (včetně diskuse použití slepé zpětné vazby) a jednak přehled historických i existujících projektů/systémů určených pro tuto úlohu. Kapitoly 4-6 jsou také teoretické, poskytují velmi dobrý přehled celé výzkumné oblasti, včetně rešerše relevantních publikací, historických i aktuálně nejlepších metod a jejich výsledků. Konkrétně se autorka věnuje metodám evaluace (Kapitola 2), problematice vyhledávání informací obecně (Kapitola 3), vyhledáváním v řečových datech (Kapitola 4), a metodám zpětné vazby (Kapitola 6). Kapitola 7 je experimentální, tvoří jádro celé práce a popisuje řadu experimentů, nových metod a jejich výsledků. Závěrem práce je kapitola osmá, ve které autorka shrnuje celou disertační práci, vyjadřuje k míře splnění jednotlivých cílů a možným směrům dalšího výzkumu.

Výsledky práce

Výsledky práce jsou několika druhů. Autorka jednak velmi pečlivě a přehledně popsala důležité metody a techniky ze zkoumané oblasti (modely pro vyhledávání informací, předzpracování dat, metody zpětné vazby, atd.), aplikovala je na testovací kolekci Malach a vyhodnotila jejich úspěšnost. Dále autorka navrhla několik modifikací existujících metod, s různým vlivem na výslednou úspěšnost vyhledávání. Nejzajímavější počin je zřejmě metoda pro stanovení optimálního počtu dokumentů pro zpětnou vazbu v závislosti na dotazu a prvotním vyhledávacím kroku. Úspěšnost této metody se sice

nepodařilo prokázat na testovací sadě témat z kolekce Malach, ale jisté zlepšení se projevilo na příbuzné úloze detekce tématu textu (této úloze je věnováno několik stran v kapitole 7, ale lze ji chápat jen jako doplňkovou, s vyhledáváním v řeči souvisí jen okrajově). Celkově nejsou výsledky nijak zvlášť převratné, ale šířka a hloubka zpracování studované problematiky je výborná, a práce tak poskytuje dobrý přehled o zkoumané oblasti a použitelnosti jednotlivých metod, což lze jistě považovat za přínos pro obor.

K práci mám několik výhrad: 1) I přesto, že od samého začátku cílí disertační práce na velmi specifickou oblast vyhledávání informací, a to vyhledávání v mluvené řeči, není toto v použitých metodách nijak využito (tedy kromě toho, že evaluace je provedena na kolekci nahrávek mluvené řeči). Mluvená řeč je automaticky přepsána do textové podoby a nadále se s ní pracuje jako s textovými daty (po rozdělení na kratší úseky, které pak odpovídají dokumentům). Metody prezentované v předkládané práci je zřejmě možné použít i pro tradiční vyhledávání v textových dokumentech a není jasné, jestli aplikace na řečová data je něčím výjimečná.
2) Od části 7.6 autorka používá v experimentech dotazy konstruované jako spojení všech tří částí popisu témat (title, description, narrative). Toto rozhodnutí je zdůvodněno v části 7.5.5 tím, že při použití kompletního popisu tématu je dosaženo lepších výsledků vyhledávání. To je pochopitelné, stejně jako fakt, že i zlepšení dosažená jednotlivými metodami jsou potom signifikantnější, ale v praxi nelze očekávat, že uživatel vyhledávacího systému bude při popisu hledané informace takto podrobný. Bylo by proto vhodnější uvádět výsledky i pro varianty dotazů, které více odpovídají realitě.
3) Autorka k evaluaci použitých metod přistupuje pouze souhrnně a prezentuje výsledky průměrované přes několik dotazů. Bylo by zajímavé znát i detailnější analýzu výsledků, např. pro kolik dotazů se vyhledávání zlepšilo a pro kolik zhoršilo, případně co eventuální zhoršení způsobilo a jestli by bylo možné mu předejít (např. jestli slova použitá pro rozšíření dotazu pomocí slepé zpětné vazby jsou vhodně vybraná apod.).

Jazyková a grafická úroveň

Práce je psaná česky, v podstatě bez pravopisných chyb a překlepů. Text je velmi dobře strukturovaný, čitelný a srozumitelný.

Publikační činnost autora

Seznam publikací Lucie Skorkovské obsahuje velmi dobrých 12 položek, většinou příspěvků na mezinárodních konferencích (zejména TSD), u 10 z nich je Lucie Skorkovská uvedena jako první autor.

Závěr

U předkládané disertační práce bych vyzvedl zejména celkovou úroveň popisu celé problematiky, množství a pečlivost provedených experimentů, analýzu jejich výsledků a v neposlední řadě také jisté vylepšení existujících metod slepé zpětné vazby. Lze konstatovat, že cíle práce byly splněny, výsledky publikovány na mezinárodních fórech, práci tudíž doporučuji k obhajobě.

RNDr. Pavel Pecina, Ph.D. 31.1. 2017, Praha

Posudek oponenta disertační práce

Název práce: Vyhledávání informací v řeči a využití slepé zpětné vazby

Autor: Ing. Lucie Skorkovská

Disertační práce je zaměřena na zlepšení metod vyhledávání informací v řeči. Konkrétně pak zapojení metod slepé zpětné vazby. Přínosem je analýza a návrh metod pro volbu optimálního počtu pseudo-relevantních dokumentů, ze kterých jsou vybírány termy pro rozšíření dotazu. V práci byly popsány a otestovány klasické metody vyhledávání, efekty předzpracování dat a varianty slepé zpětné vazby. Univerzálnost poznatků byla zkoumána také při detekci témat v textu.

Celkově jsou řešená témata aktuální, práce dává velmi pěkný přehled problematiky vyhledávání v řeči a experimenty znamenají přínos pro vyhledávání v českých nahrávkách. Cíle práce byly tedy splněny.

Použité metody a postupy jsou srozumitelně popsány. Formálně je práce v pořádku, až na kvalitu a čitelnost některých obrázků (např. 6.6). Práce je dobře strukturovaná a vyskytuje se v ní jen poměrně malé množství překlepů.

Publikační činnost je dostatečná. Počet recenzovaných publikací vydaných kvalitními mezinárodními vydavatelstvími je nadprůměrný (12 článků). Navíc autorka je ve většině případů prvním autorem. Jedná se o konferenční příspěvky převážně lokálního, popř. evropského, významu. Postrádám publikaci s celosvětovým dosahem.

K práci mám následující připomínky:

- State-of-the-art: Citovaných publikací je velké množství. Práce svědčí o dobré orientaci v problematice, ale absence citací z posledních let (<10 za posledních 5 let) budí pocit, že se na problémech už tolik nepracuje. Z popisu historie a aplikací lze také zjistit velmi málo o aktuálním dění. Více informací o MediaEval by to možná osvětlilo.
- Hodnotící metody: v přehledu bych doplnil Discounted cumulative gain.
- Odstranění stop-slov bylo diskutováno v literatuře hojně. Chybí reference. Odstranění stop-slov pro další experimenty v práci je diskutabilní, vyhledávače je ve většině případů používají. Jaká je časová/paměťová úspora?
- Česká kolekce spontánní řeči: pokud se nejedná o unikátní experimenty s vyhledáváním v kolekci CLEF CL-SR, lze dosažené výsledky srovnat s výsledky v literatuře?
- Dají se nějak vysvětlit velké rozdíly v úspěšnosti vyhledávání trénovacích a testovacích témat?
- Naive Bayes je ve většině případů překonáván jinými klasifikátory. Otázkou je, zda by v případě použití lepšího klasifikátoru zůstaly závěry stejné.
- Detekce tématu: kvalitu metody by bylo dobré také otestovat na standardní anglické kolekci. Experimenty by pak mohly mít větší impakt.
- Obrázek 6.1 a 6.3 je totožný, znázornění uživatele by mohlo ukázat rozdíl.

Detaily:

- Str. 29: $k_3=7$ -> $k_2=7$
- Str. 29: V ve vzorci 4.17

Disertační práce prokazuje předpoklady autorky k samostatné tvořivé práci a doporučuji ji k obhajobě.

V Plzni, 24. ledna 2017

Doc. Ing. Josef Steinberger, Ph.D.