Posudek oponenta diplomové práce

Autor/autorka práce: **Petr Matejovič**

Název práce: **Rozšiřování dotazu pomocí sémantických prostorů**

Obsah práce

Práce se zabývá úlohou rozšiřování dotazů za pomocí aplikace sémantických prostorů (LSA, COALS), což je velmi zajímavá a v dnešní době aktuální úloha. Kvalita rozšiřování dotazů je měřena v oblasti IR (Information Retrieval). Na úvod je třeba říci, že si diplomant vybral poměrně hodně náročné téma, které od něj vyžadovalo pochopení mnoha metod v různých oblastech.

K samotnému obsahu práce mám řadu výhrad:

Str. 10: Chybně nadefinované míry Precision a Recall. RET a REL jsou množiny dokumentů a ne počty dokumentů.

Str. 14: Diplomant tvrdí, že česká data nejsou v korpusu CLEF dostupná. Omyl, je zde spousta jazyků včetně češtiny.

Chybně popsaná lemmatizace. Lemmatizace vybírá správné lemma nejenom na základě příslušného slovního tvaru, ale i kontextu, ve kterém se nachází (tzn., řeší i nejednoznačnost slov).

Str. 15: Kapitola 2.5.2 první odstavec - úplně špatně.

Str. 18. Kosinová míra je kosinus úhlu, ne úhel sám.

Str. 21: Kapitola 2.6. Špatná definice rozšiřování dotazu, plus nepochopitelný příklad.

Str. 23: Architektura vyhledávání s použitím rozšiřování dotazů by měla být popsána úplně jinde (ne v kapitole o korpusu CLEF).

Str. 27: Špatně nadefinovaná míra MAP (Mean Average Precision). Q je počet dotazů. Chybí definice AP (Average Precision). Navíc, tato definice by měla být v teoretické části a ne v analýze.

Str. 44: Proč bylo při testování používáno zrovna 200 nejlepších výsledků? Není zdůvodněno. Používají se nestandartní míry pro vektorový vyhledávací model. Precision a Recall nezohledňují pořadí nalezeného dokumentu, proto se standardně používá zejména MAP, R-prec a P@10 (Precision při velmi malém množství vrácených dokumentů, typicky 10).

Str. 47 tabulka 5-3. Jediné výsledky, které dávají smysl. Při rozšíření dotazů o 1 variantu došlo k mírnému zvýšení MAP. Ve všech ostatních případech, je MAP velmi výrazně zhoršena při použití rozšiřování dotazů.

Str. 51: Kapitola 5.2.3 je celá naprosto zbytečná. To, že při zvyšování počtu vrácených dokumentů se bude zmenšovat Precision a narůstat Recall je jasné.

Str. 52: Kapitola 5.2.4 je zbytečná. Testování lemmatizátoru nemá s cíli práce nic společného.

Jelikož veškeré experimenty v práci jsou měřeny vždy pouze na 50 dotazech, je nutné k výsledkům připojit také test statistické významnosti. Pro tuto práci by byl vhodný například párový t-test pro porovnání dvou metod (popsaný například v [Hull, 1993]). V práci ale tyto výsledky nejsou.

V části o výsledcích experimentů mi chybí statistiky korpusu (počet slov, počet tokenů, počet dokumentů). Dále, tabulky s výsledky nejsou popsané. Není vysvětleno, co znamenají zkratky v hlavičce tabulek.

Závěr práce poněkud neodpovídá dosaženým výsledkům. Autor sice prezentuje zvýšení Precision a Recall, ale tyto míry nejsou vhodné pro vektorový vyhledávací model, který zohledňuje pořadí výsledků. Naopak, míra MAP, která je pro danou oblast považována za standard, byla ve většině případů výrazně zhoršena.

Autor zdůvodňuje snížení MAP použitím normalizačních metod zpracování textu. Za prvé, rozšiřování dotazů není normalizační metoda. Za druhé, v jediném testu kde byla použita normalizační metoda (lemmatizace – kapitola 5.2.4) se MAP mírně zvýšila. Což je tedy v přímém rozporu.

Celý proces testování na mě působí velmi chaoticky. Navíc se zdá, že autor se zaměřoval na věci, které nejsou pro práci příliš důležité. Například vytvořil přehledné webové rozhraní (s řadou funkcí) pro testování rozšiřování dotazů.

Formální úroveň

I přesto, že je dokument vytvořen v MS Word, je úprava textu obstojná. Vyjadřování v českém jazyce je dobré a z hlediska formální stránky nemám k práci žádné výhrady.

Práce s literaturou

Autor v práci cituje převážně zahraniční časopisy a další hodnotou literaturu. Diplomantova práce s literaturou je v pořádku až na dvě malé výjimky: Str. 19: Citace u algoritmu HAL je uvedena až uprostřed textu (ne na začátku, kde by se čekalo). Literatura není řazena podle abecedy.

Splnění zadání

Z práce je patrné, že diplomant odvedl velké množství práce a vyzkoušel velké množství metod pro rozšiřování dotazů. Tato práce je však velmi chaoticky a nepřesně popsána. I přes řadu chyb a nepřesností, které jsem zmínil, byly cíle práce více méně splněny.

Dotazy k práci:

1) Rozšiřování dotazů ve Vaší práci vedlo skoro ve všech případech k velmi výraznému snížení míry MAP (mean average precision), která je považována za standartní míru. V práci však prezentujete zvýšení míry Precision a Recall. Čím si toto vysvětlujete?
2) Čím si vysvětlujete, že Vámi navržené metody spíše nefungují? Málo dat pro trénování sémantických prostorů? Špatné nastavení sémantických prostorů?
3) Jestliže jste se chtěl v práci zaměřovat na měření Precision a Recall, proč jste nepoužil booleovský model pro vyhledávání namísto vektorového modelu, kde je důležitá zejména MAP?

Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni 29.8.2013

Ing. Tomáš Brychcín

Literatura

Hull, D., 1993. *Using statistical testing in the evaluation of retrieval experiments.* In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '93. ACM, New York, NY, USA, pp. 329-338.

SOUHLASÍ S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky