Oponentský posudok dizertačnej práce

Meno doktoranda: Ing. Jan Zelený

Názov práce: Web page segmentation utilizing clustering techniques

Školiteľ: Doc. Ing. Jaroslav Zendulka, CSc.

Predložená dizertačná práca sa zaoberá témou segmentácie webových stránok na základe zhlukovania.

Téma dizertačnej práce je aktuálna a považujem ju za výsostne relevantnú z pohľadu oblasti extrakcie informácií na webe, ktorá má dôležité implikácie aj do iných oblastí s perspektívou praktickej aplikácie, čo považujem za obzvlášť cenné.

Práca je rozdelená do troch logických častí. V prvej, ktoré sú pokryté kapitolami 2 až 4, doktorand opisuje stav v oblasti, v druhej (kapitoly 5 až 7) predstavuje návrh metód pre segmentovanie webových stránok. V tretej časti reprezentovanej kapitolami 8 a 9 opisuje implementáciu a overenie.

Analýza problémovej oblasti a prehľad aktuálneho stavu poznania patrí k slabším častiam práce. Vovedenie do problematiky nie je ideálne, hneď zo začiatku je miestami zmätočné (napr. zaradenie information retrieval ako úlohy data mining je bez ďalšieho kontextu prinajmenšom diskutabilné, s. 3) a predpokladá mnohé znalosti čitateľa, ktoré sú rozvité až v neskorších častiach práce, čo komplikuje úvodné pochopenie motivácie/cieľov práce. Chýba skoré vysvetlenie alebo presnejšie vymedzenie kľúčových doménových pojmov (napr. čo je to šablóna (template), blok (block) v kontexte tejto práce, a pod.). Chýba tiež lepšie ukotvenie východísk, napr. v podobe nejakej deskriptívnej charakteristiky dnešného Webu – počet stránok a sídel, typy sídel, a pod., ktoré by podložili uvedené tvrdenia o šírke aplikácii v rámci dolovania v dátach, či o potrebe škálovateľnosti. Pre aký typ obsahu má navrhovaná metóda partikulárny zmysel?

Stav poznania považujem za pomerne dobre zmapovaný, avšak často nejde do dostatočnej hĺbky, čo je dané nezriedkavou nejasnosťou (napr. tf-idf predstavené ako „distance metrics“) alebo primalým rozsahom opisu (napr. opis prác na str. 20). Len zriedka sa autor vyjadruje k presnej úspešnosti/výkonnosti analyzovaných metód, nediskutuje metodiku overenia ani dáta, na ktorých boli metódy overované, a vzájomne ich neporovnáva. Dôsledkom je, že analytická časť práce nedáva ucelený kritický pohľad na stav poznania, skôr len vymenúva jednotlivé metódy a prístupy (čo v konečnom dôsledku komplikuje posúdenie originálneho prínosu autora a zasadenie neskoršieho návrhu do kontextu existujúcich riešení). Oceňujem klasifikáciu do skupín prístupov (code-based, vision-based). Pokrytie opísaných prístupov v problémovej oblasti je dobré, nie však vyčerpávajúce. Chýba mi aspoň zmienka o prístupoch týkajúcich sa extrakcie hlavného textu z webových stránok (napr. [1, 2] – referencie na konci posudku), ktoré považujem v určitom zmysle za príbuzné, a pre skupinu text-based metód dokonca relevantné.

Aj napriek uvedeným výhradám a faktu, že niekoľko ťažiskových referencovaných prác je skôr starších, analýza celkovo prináša dobrý základný prehľad o stave poznania ako podklad pre návrh vlastného riešenia – metód súvisiacich so segmentovaním webových stránok.

Celkový návrh metód považujem, v kontraste s predošlou časťou práce, za veľmi dobre rozpracovaný a premyslený. Autor dôkladne a podrobne opisuje jednotlivé kroky návrhu riešenia, ktoré v podstate pozostáva z dvoch častí: zhlukovania častí webovej stránky (boxes) a porovnávania predlôh (templates). Uvedený návrh hodnotím ako cenný pôvodný prínos

dizertačnej práce, ktorý možno považovať za rozšírenie stavu poznania v oblasti. Vytknúť sa dajú len občasné nejasnosti (napr. s.36: čo presne znamená modifikácia komponentu o x %, s. 38: aké „experimental observation“ má autor na mysli? (žiada sa tvrdenie lepšie podložiť); a pod.), ktoré však nie sú zásadného charakteru. Špeciálne vyzdvihujem úsilie vynaložené na formalizáciu problému, ktorý je reflektovaný do 28 definícií a 5 formalizovaných algoritmov. Úroveň rozpracovania problematiky je naozaj úctyhodná. Oceňujem, že celkový návrh nie je len sofistikovanou teoretickou metódou, ale zohľadňuje aj aspekty časovej a priestorovej zložitosti.

Slabšou stránkou tejto časti práci je chýbajúca sumarizujúca diskusia k celkovému návrhu metód, najmä zhodnotenie ich možných ohraničení. Okrem toho by inak veľmi dobre rozpracovanému formálnemu opisu metód pomohlo viac príkladov či ilustrácií (napr. v časti 7.2 týkajúcej sa mapovaní).

V rámci ďalšej časti práce autor zdôvodňuje implementačné rozhodnutia a špecifiká realizácie navrhnutého prístupu.

Overenie navrhnutej metódy sa javí skôr povrchné. Rozsah je relatívne krátky (8 z celkových 74 strán), čoho dôsledkom je, že chýba viacero dôležitých informácií: explicitná formulácia hypotézy alebo výskumných otázok, opis metodológie overenia či deskriptívna charakteristika použitých dát. Zostavenie zlatého štandardu pre overenie (referenčnú segmentáciu) je opísaný len veľmi stručne. Nie je napr. jasné, aká bola zhoda troch spomínaných anotátorov (zaujímavé by bolo vyhodnotiť ju explicitne kvôli identifikácii mantinelov úspešnosti overovanej metódy).

Prezentované výsledky overenia ukazujú výkonnosť metódy v porovnaní s metódou VIPS z r. 2003. V oblasti vyhodnotenie časovej efektivity preto výsledky nevyznievajú prekvapujúco. Naopak, čo sa týka úspešnosti, vo všeobecnosti lepšie výsledky sú dosiahnuté starším riešením. Uznávam náročnosť celkového overovania tohto typu metód, očakával by som však lepšie rozpracovanie celkového prístupu spolu s identifikáciou aspektov, v ktorých navrhovaná metóda oproti VIPS vyniká. Chýba mi tiež diskusia k interpretácii hodnôt finálneho F-skóre. (Možno ich vzhľadom na existujúce ohraničenia považovať za dobré? Za zlé?) Celkovo hodnotím overenie skôr za veľmi základné, napriek tomu však oceňujem prítomnosť kvantifikácie výkonu metód (z pohľadu úspešnosti aj časovej efektivity).

Prácu uzatvára zhodnotenie výsledkov a veľmi stručný náznak možných vylepšení a budúcej práce.

Z formálneho hľadiska je práca na veľmi dobrej úrovni, obsahuje len menej závažné nedostatky.

Celkové zhodnotenie

Doktorand sa v rámci práce venoval aktuálnej a relevantnej téme segmentácie webových stránok na úrovni zodpovedajúcej aktuálnemu stavu poznania v odbore. Najcennejším prínosom práce je dôkladne rozpracovaný návrh vizuálne-orientovanej metódy (metód), ktorý možno považovať za originálny prínos autora. Slabšími časťami práce sú spracovanie aktuálneho stavu poznania a overenie návrhu metód. Jadro dizertačnej práce bolo publikované v podobe piatich vedeckých príspevkov, čo považujem za dostatočnú úroveň. Doktoranda ocenila odborná komunita prijatím príspevkov aj v karentovaných časopisoch, čo považujem za uznanie dostatočnej vedeckej erudície.

Na základe predloženej práce konštatujem, že doktorand preukázal schopnosť výskumnej práce. Aj napriek niekoľkým výhradám dizertačná práca podľa môjho názoru zodpovedá všeobecne uznávaným požiadavkám k udeleniu akademického titulu PhD.

Referencie v texte:

[1] Fu, L., Meng, Y., Xia, Y., & Yu, H. (2010, July). Web content extraction based on webpage layout analysis. In *Information Technology and Computer Science (ITCS), 2010 Second International Conference on* (pp. 40-43). IEEE.

[2] Sun, F., Song, D., & Liao, L. (2011, July). Dom based content extraction via text density. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 245-254). ACM.

V Bratislave, 8. 6. 2017

_____

doc. Ing. Marián Šimko, PhD.