Oponentní posudek na disertační práci uchazeče

Michal Nykl

Oponent: Miroslav Šnorek

a) Zhodnocení významu disertační práce pro obor

Hodnocení významnosti prvků v množině se používá, používalo a používat bude v nejrůznějších disciplínách. Má obvykle velký význam, protože z podstaty věci bývá zapotřebí nějaké entity srovnávat a stanovovat jejich pořadí. Jen někdy bývá řešení této úlohy snadné. V řadě případů totiž pro stanovení pořadí neexistují objektivní kritéria a postupuje se podle nějakých (obecně) přijatých zásad, objektivizovaných metodik. Různé metodiky potom, přirozeně, vedou k různým výsledkům, různému pořadí, a vyvstává otázka: které pořadí je to správné?

Disertant si vybral v akademickém světě trvale aktuální úlohu - hodnocení autorů vědeckých publikací - bibliometrii. Téma je z různých objektivních důvodů (například výběr nejkvalifikovanějšího zpracovatele posudku grantového, nebo podobného posudku) aktuální a žádané. Druhá úloha, kterou disertant v rámci předložené práce řešil, je z oblasti zpracování textových dokumentů. I ta je v současnosti aktuální. Pro obě úlohy autor nabízí jím navržené varianty metody (Pagerank).

b) Vyjádření k postupu řešení problému, použitým metodám a splnění určeného cíle

Autor se řešenému problému věnuje dlouhodobě. Nepřekvapuje tedy, že dobře zná state-of-the-art disciplíny (kapitola 2 disertace, přiměřeně rozsáhlý seznam literatury). Upozorňuje, že volbou dobré metodiky, ale nevhodnou interpretací výsledků, lze hodnocení, např. autorů, znehodnotit (populární vs. prestižní). Hlavní autorovou inspirací je metoda Pagerank, kterou vylepšil.

Cíle disertace explicitně a přehledně autor formuloval v odst. 1.1, stav jejich dosažení zase v odst. 7.2. Tyto dva zásadní odstavce mi nakonec pomohly se ve struktuře práce vyznat. Přestože jsem měl pocit, že se práce dobře čte, snaha najít v textu disertantovy metody dlouho končila ve slepých uličkách. Ztrácel jsem se v použité symbolice, v použitých datových kolekcích (CiteSeer, DBLP a WoS) a v popisu provedených experimentů včetně vyhodnocení závěrů z nich (kapitoly 3 až 5). Jsem si ale jist, že uvedená výhrada je (zejména) mým problémem, protože autorova strukturace je nutná a bezchybná. Nakonec rád uvádím, že ve srovnání vlastních a konkurenčních výsledků, ve vyvozování závěrů z těchto srovnání a v jejich interpretaci vidím autorský přínos disertanta.

Pro metody použité v bibliometrické úloze (hodnocení autorů) je zřejmě zásadní Obr. 5.4. Ten shrnuje celou škálu experimentů s metodami disertantem použitými a poněkud vysvětluje použitou symboliku.

Cíle disertace jsou dle mého názoru splněny.

Sympatické mi bylo, že správnost pořadí stanoveného svými metodami prověřil tím, že se pokusil predikovat laureáty odborných ocenění (Cena ACM). Konstatuji, že predikoval úspěšně (tabulky Tab. 5.5 a Tab. 5.6).

c) Stanovisko k výsledkům disertační práce a k původnímu konkrétnímu přínosu disertanta

Za původní přínos považuji (ve shodě s disertantem) ty uvedené v kapitole 7.2 práce. Snad se nemýlím v odhadu, že se dozajista přínosné závěry této práce se nakonec dostanou na patřičná místa alespoň v tuzemsku prostřednictvím třeba vedoucího práce, který (doufám) v předmětné oblasti na nejvyšší národní úrovni pracoval/pracuje.

d) Vyjádření k systematice, přehlednosti, formální úpravě a jazykové úrovni disertační práce

Práce obsahuje minimum překlepů nebo chyb. Vzhledem k tomu, že téma je celosvětově aktuální, vidím jako její významný nedostatek, že je publikována česky.

Po prostudování předloženého textu konstatuji, že mě uchazeč přesvědčil o schopnosti shromáždit k danému tématu velký objem relevantních informací, které dokáže analyzovat, kriticky zhodnotit a z nich vyvozovat odůvodněné závěry. Dovede používat přiměřený matematický aparát i symboliku.

e) Vyjádření k publikacím studenta

Nejkvalitnějšími publikacemi autora jsou ty v časopise Journal of Informetrics, Elsevier, v letech 2014 a 2015 (IF=2,412). Považuji je rozhodně za publikace odpovídající úrovni posuzované vědecké kompetence autora. I počet citací, podle mého názoru, odpovídá vědecké kompetenci autora.

f) Jednoznačné vyjádření oponenta

Předloženou disertační práci k obhajobě jednoznačně

doporučuji.

V Praze, 20.1.2016

doc. Ing. Miroslav Šnorek, CSc.

Oponentský posudek disertační práce

Kandidát: Ing. Michal Nykl
Název práce: Hodnocení významnosti variantami PageRanku
Oponent: Ing. Tomáš Kliegr, Ph.D., FIS VŠE

Popis práce

Předložená disertační práce se zabývá metodami pro hodnocení významnosti autorů vědeckých publikací. Z algoritmického hlediska autor používá metodu PageRank včetně personalizovaného rozšíření. Autor tuto metodu přizpůsobuje pro použití na citační sítě. Práce obsahuje řadu experimentů, m.j. s cílem nalézt nejvhodnější metodu pro stanovení hodnot uzlů v citační síti, která vstupuje do algoritmu PageRank. Navržené metody jsou porovnány se základními algoritmy, které hodnotí autory přímo na základě počtu citací. Evaluace je provedena na několika rozsáhlých datasetech (DBLP, Citeseer, ISI Web of Science). Dále práce obsahuje jednu kapitolu, kde je představena nová metoda pro obohacení dokumentů o klíčová slova s využitím propojených dat a algoritmu PageRank, přičemž evaluace je provedena v doméně klasifikace dokumentů.

Hodnocení práce

Hlavním přínosem dizertační práce je řada poznatků v oblasti bibliometrie. Uvedená zjištění jsou doložena řadou experimentů a podrobnou diskuzí výsledků. Výzkumné otázky jsou jasně formulovány. Postup jejich řešení považuji za adekvátní a odpovídající dobré praxi v oboru. Práce obsahuje podrobný úvod do problematiky, rešerši literatury, část představující použité algoritmické řešení, a rozsáhlou evaluaci. Závěr práce shrnuje dosažené výsledky a pokládá je do kontrastu s cíli práce.

Práce je přehledně strukturována, obsahuje řadu tabulek a obrázků, které výrazně přispívají k čitelnosti textu. Poněkud nekoncepční je sekce 2.4.1, která současně představuje existující algoritmus PageRank tak i jeho rozšíření představené autory. Práce neobsahuje zcela jednoznačné srovnání navržených metod se state-of-the-art algoritmy řešícími stejný problém, a to ani v rešeršní ani evaluační rovině. Absenci evaluace s jinými přístupy přitom autor vytýká jiným algoritmům (Eigenfactor score na str. 69), přitom ale samotná práce neobsahuje experimentální evaluaci zahrnující algoritmy navržené jinými autory.

Jistý nedostatek vidím v tom, že experimentální evaluace je sice rozsáhlá, ale je provedena pouze v několika věcných kategoriích rejstříku WOS (počítačové vědy). Citační zvyklosti se výrazně liší v jednotlivých vědních disciplínách. Artikulované závěry také zcela ignorují vlastní výsledky dosažené autorem na kolekcích Citeseer a DBLP. Konkrétně jeden z hlavních výsledků práce - *zjištění, že metody, které pracují s citační sítí publikací, poskytují dokonalejší hodnocení autorů, než metody pracující s citační sítí autorů* – byl prokázán pouze na kolekci WOS, na zbylých dvou kolekcích byl závěr opačný (str. 41, 44). Další bibliografické databáze byly vynechány již v rešeršní části. V sekci 2.2.2 autor uvádí několik specializovaných databází pro informatiku (CiteSeer, DBLP,ArXiv), přičemž opomíjí významné databáze pro jiné vědní oblasti (např. pubmed pro biomedicínu).

Jako málo zdůvodněný se mi jeví závěr autora upřednostňující komerční dataset WOS před volně dostupnými datasety Citeseer a DBLP. Ve svých analýzách autor použil verzi DBLP z roku 2004, přičemž konstatuje, že tato databáze je nekompletní, protože obsahuje velký počet slepých a izolovaných vrcholů (str. 33). Současně ale uvádí, že "kolekce [DBLP] z let 2006 a 2009 obsahují téměř totožný počet citací jako verze z roku 2004, ale daleko více publikací a autorů" (str. 38).

Kapitola 6 popisuje potenciálně zajímavý přístup navržený autory pro klasifikaci dokumentů s využitím algoritmu PageRank a propojených dat. Nicméně s ohledem na absenci rozsáhlejší evaluace (chybí srovnání s výsledky jiných algoritmů) a absenci jakéhokoliv srovnání navržené metody s existujícími příbuznými přístupy, představuje tato kapitola odbočku od hlavního tématu, a pro celkové hodnocení práce jsem ji neuvažoval.

Jazyk a formální úprava jsou adekvátní dizertační práci. Je obsaženo malé množství překlepů a odborného žargonu (např. pojem "Baseliny" na str. 68). Autor také někdy používá anglický pojem, i když existuje ustálený český ekvivalent ("in-degree"). Některé důležité pojmy nejsou jednoznačně definovány ("hodnota publikace").

Podrobné připomínky k textu práce jsou uvedeny v příloze.

Otázka na autora

Na str. 20 autor uvádí: "Naší úpravou v prezentovaných vzorcích, které popisují výpočet PageRanku pro jeden prvek, bylo přidání části s ošetřením slepých vrcholů. Tato část umožňující urychlení výpočtu nebyla v původních pracích (Brin a Page 1998; Page et al. 1999; Langville a Meyer 2006) použita."

Z textu mi není jasné, k jaké části vzorce se tato věta vztahuje. Úsporné ošetření slepých vrcholů zajišťuje člen *a* v rovnici 5.3.1 na str. 50 v LANGVILLE, Amy N. a MEYER, Carl D., 2006. Google's PageRank and Beyond The Science of Search Engine Rankings. Princeton, NJ, USA: Princeton University Press. ISBN 9780691152660.

Závěr

Výsledky úzce související s předkládanou prací byly publikovány v impaktových časopisech a sbornících indexovaných ve významných citačních rejstřících, publikační činnost kandidáta tak považuji za přiměřenou.

I přes určité výše uvedené výhrady k zobecnitelnosti některých dílčích výsledků mimo oblast počítačových věd a citační rejstřík WOS, považuji původní konkrétní přínos předkladatele za adekvátní a cíle práce za splněné (s platností částečně omezenou pro WOS a počítačové vědy). Výsledky výzkumu považuji za využitelné v daném oboru - bibliometrii. K velikosti přínosu se nedokáži vyjádřit, protože práce neobsahuje benchmark s algoritmy jiných autorů.

Domnívám se, že předložená práce po obsahové i formální stránce *splňuje požadavky pro udělení akademického titulu Ph.D.*.

Práci doporučuji k obhajobě.

21.1.2016, Darmstadt

Ing. Tomáš Kliegr, Ph.D.

Příloha: podrobné připomínky k textu práce

str.14:

"Výhodou ħ-indexu je, že produktivnějším autorům penalizuje spolupráci se začínajícími autory." - Z textu není jasné, proč penalizaci spolupráce se začínajícími autory autor považuje za výhodu.

"g z jeho top článků" - nejasný význam slova top

str. 15,16
Jméno míry by se mělo přeložit úplně nebo vůbec ("Degree centralita", "betweenness centralita"). Pojmy degree, in-degree a out-degree mají české ekvivalenty (vstupní a výstupní stupeň vrcholu). Autor si tohoto je vědom (3. odstavec str. 15, předposlední odtavec str. 16), proč tedy nepoužít české pojmy, nebo alespoň nestandardní terminologii neodůvodnit poznámkou?

str. 17
V textu je uvedeno, že vztah (2.5) je základní vzorec PageRanku dle Page et al. (1999). V tomto článku nicméně přesně tento vzorec není uvedený, nejblíže je mu vztah uvedený v definici 1 v Page et al. (1999), který je označen jako "zjednodušený" (což je něco jiného než "základní").

V sekci 2.4.1 je použit pojem "citační prestiž vrcholu" s odkazem na článek Page et al. (1999), tam se ale slovo prestiž vůbec nevyskytuje.

str. 18
Není zřejmé, odkud je převzat seznam třech možných způsobů řešení problému "slepých vrcholů". Mezi uvedenými metodami není zahrnuto odebrání "slepých odkazů" (dangling links) z grafu a vrácení po ukončení výpočtu, tak jak je navrženo v sekci 2.7 publikace

PAGE, Lawrence, BRIN, Sergey, MOTWANI, Rajeev a WINOGRAD, Terry, 1999. The PageRank Citation Ranking:
Bringing Order to the Web. 1999. Stanford: Stanford InfoLab. [vid. 18. prosinec 2012].
Dostupné z: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

"Při výpočtu PageRanku využitím matematického zápisu výpočtu pro jeden prvek nemusíme výstupní hrany ze slepých vrcholů doplňovat přímo do grafu, ale stačí s nimi pouze počítat, jak ukazuje námi navržený vzorec"
Zde by možná bylo vhodné uvést, že podobná optimalizace výpočtu je uvedena v knize Langville & Mayer, 2006.

"Analýzou chování reálných uživatelů Webu autoři zjistili, že teleport uživatelé využívají průměrně jednou za 7 kroků."

Toto tvrzení se mi nepodařilo najít. v odkazované publikaci: BRIN, Sergey a PAGE, Lawrence, 1998. The anatomy of a large-scale hypertextual Web search engine. Computer
Networks and ISDN Systems. 30(1-7), 107–117. ISSN 01697552.
Dostupné z: doi:10.1016/S0169-7552(98)00110-X

"Přesto se ale obvykle před iterací nastavují hodnoty vrcholů ... nejlépe na hodnoty blízké konečnému výsledku."

Chybí zdroj k tomuto tvrzení

str. 20
"Pojem personalizace zavedli Page et al. (1999)," - vhodnější by bylo formulovat jako "Pojem personalizace v kontextu ..."

V sekci není uvedeno, na základě jakého výpočtu se získá upravená (personalizovaná) matice G (H).

U vzorců 2.11-2.14 chybí zdroj. V citované publikaci Page et al. 1999 nejsou uvedeny.

str. 21
"Využitím faktoru tlumení d=0,55 autoři vkládají popularitu do výpočtu PageRanku hodnotícího spoluautorství."

To je poněkud zkratkovitá formulace, jejíž účel je nejasný v tomto kontextu. V uvedeném článku autoři evaluují několik hodnot damping faktoru, pro hodnotu 0.55 pak konstatují, že "the coauthorship network topology and author's citation counts have nearly the same impact towards the PR_W score".

str.26

Vzorec 2.30: chybí popis proměnné ai a vysvětlení členu 0.01.

str. 37

Pojem "hodnota publikace" není definován. Je tím myšlena hodnota vypočítaná algoritmem PageRank? Jedná se o klíčový pojem v dizertační práci, který by mohl být odlišen pro lepší čitelnost např. kurzívou.

str. 39
Hodnoty s sloupci Slepé a Izolované v tabulce 3.2 jsou překvapivě vysoké. Podle popisu hodnota Slepé se vztahuje k vrcholům, které nikoho necitují, a Izolované jsou vrcholy, které nejsou nikým citovány a jsou slepé. Zde to patrně znamená, že citující resp. citovaná publikace není zahrnuta v DBLP (Citeseer), protože je prakticky vyloučeno, aby akademická publikace necitovala alespoň jednu jinou publikaci.

Situace, kdy je většina vrcholů v kategorii Slepé se mi jeví jako nepřirozená pro většinu domén a bibliografických databází. Z poznámky obhajující staré použité verze DBLP vyplývá, že pokud by byl použit novější dump těcho bibliografických databází obsahující "daleko více publikací a autorů", tak by k tomutu jevu nemuselo dojít.

Na str. 45 Autor nejednoznačnost výsledků přisuzuje nekvalitní databázi s malým počtem indexovaných citací a odůvodňuje tím použití ISI WoS. Jak je uvedeno výše, řešením by mohlo být též použití novějších dumpů.

str. 45
"Z analýzy této metody jsme usoudili, že držitelé Coddovy ceny měli značný vliv na výběr článků indexovaných v DBLP a na kvalitu indexace jejich citací." Autor je tedy toho názoru, že držitelé Coddovy ceny se podílejí na výběru indexovaných časopisů a konferencí v DBLP, přičemž

upřednostňují ty publikace, které je citují? Nemůže existovat i jiné vysvětlení, například to, že Coddova cena nejlépe odpovídá tématickému zaměření DBLP?

str. 46.
Použití "seznamu vysoce citovaných výzkumníků ISI" by si zasloužilo podrobnější zdůvodnění, protože tato metrika je odvozena od počtu citaci v ISI, který úzce koreluje s výzkumnou otázkou - zda je "prestiž je lepší mírou pro hodnocení autorů než popularita", přičemž "Popularita je zastoupena počítáním citací." Tato připomínka je zmíněna a uznána autory na straně 58, měla by se ale objevit dříve, a to zejm. při hodnocení výsledků v kapitole 4.

str. 47
"Z toho důvodu jsme předpokládali, že hodnocení autorů na základě vyhodnocení citační sítě publikací poskytne lepší pořadí autorů, než hodnocení autorů na základě vyhodnocení citační sítě autorů. Přestože v kolekcích CiteSeer a DBLP se tento předpoklad nepotvrdil, tak zde popsané experimenty s kolekcí WoS ukázaly, že byl správný."

Z uvedené věty vyplývá, že na dvou ze třech evaluačních databází neposkytlo vyhodnocení citační sítě publikací poskytne lepší pořadí autorů a na jedné ano. Z toho nelze dle mého názoru obecně odvodit, že hodnocení autorů na základě vyhodnocení citační sítě publikací je vhodnější než hodnocení autorů na základě vyhodnocení citační sítě autorů.

str. 48
U závěru "kolekce WoS obsahuje z pohledu počtu indexovaných citací kvalitnější záznamy než kolekce DBLP." autor neupozorňuje na to, že je srovnávána verze DBLP z roku 2004 s ISI WoS z roku 2009.

citace

- u dokumentu Page et al 1999 je špatně uveden rok, má být 1999

str. 57
"Naším cílem bylo zjistit, zda má smysl při hodnocení zvýhodňovat autory, kteří jsou ve výčtu autorů publikace na předních pozicích."

V computer science je poměrně obvyklé, že vedoucí autorského kolektivu je na poslední pozici. Zvážili jste i tuto hypotézu? V některých vědních disciplínách je obvyklé autory uvádět abecedně (http://mathoverflow.net/questions/19987/math-paper-authors-order), což omezuje aplikovatelnost výsledků experimentů s pořadím autorů. Tento problém není v práci diskutován.

str. 70
Hodnota VAL(Q) je vysvětlena jako hodnota publikace Q, nicméně pojem "hodnota publikace" není dále definován.

str. 77
Jak byla stanovena optimální hodnota d=0.85? Na str. 65 je uvedeno, že autoři testovali d z intervalu [0,15;0.85]. Pokud hodnota 0.85 poskytla nejlepší výsledky, měl být tento interval rozšířen.

str. 86
"Který klasifikaci obohatil o sémantické informace získané PageRankem z Linked Data".

Použití pojmu "sémantická informace" pro hodnotu PageRanku se mi nezdá v kontextu výzkumu v oblasti sémantického webu vhodné. Celkově úvod k této kapitole není příliš přesně napsán.

str. 87
"My za zdroj Linked Data prezentovaných v této kapitole zvolili DBpedii, což je sémanticky obohacená Wikipedie". Toto není zcela správná definice, viz http://wiki.dbpedia.org/.

str. 88
Zaužívaný český pojem pro Linked data je "propojená data".

str. 90
Není jasné, co myslí autor pod použitím chi kvadrát testu pro výběr klíčových slov.

str. 94
Je použit Rocchio klasifikátor, který ale nebyl popsán. Jelikož se v předpokládám jedná o výpočet kosinové míry podobnosti nad BOW reprezentací, autoři ho mohli pro úplnost v textu uvést. Algoritmus v citovaném článku Rocchio 1971 (str. 87) byl navžen pro relevance feedback, nikoliv pro přímočarou klasifikaci dokumentů.

str. 95
Na obrázku 6.6 je algoritmus Rocchio-LD, který není zmíněn v textu.

str. 96
Závěr je nepřiměřeně stručný, chybí srovnání s jinými algoritmy a celkově porovnání navržené metody s existujícími přístupy.