Moderné informetrické metódy hodnotenia vedeckého výskumu

Dalibor FIALA[1], Martin DOSTAL[1], Ján PARALIČ[2],
Gabriel TUTOKY[2], Cecília HAVRILOVÁ[2]

[1]*Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni*
Univerzitní 2732/8, 30614 Plzeň
dalfia@kiv.zcu.cz, madostal@kiv.zcu.cz
[2]*Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach*
Letná 0, 042 00 Košice
jan.paralic@tuke.sk, gabriel.tutoky@tuke.sk,
cecilia.havrilova@tuke.sk

Abstrakt. Hlavným cieľom tohto príspevku je informovať o bilaterálnom Česko-Slovenskom výskumnom projekte zameranom na analýzu súčasných, ako aj návrh a overenie nových scientometrických ukazovateľov, vychádzajúcich z metód analýzy citačných sietí a metód dolovania znalostí z textov. V rámci metód založených na analýze sietí je pritom hlavná pozornosť venovaná možnostiam adaptácie algoritmu PageRank pre potreby uvedeného cieľa. V rámci metód založených na použití dolovania znalostí z textov ide predovšetkým o modelovanie vzťahu medzi scientometrickými ukazovateľmi významnosti publikácií a ich atribútmi získanými metódami dolovania v textoch.

Klíčová slova: citačné siete, scientometrické ukazovatele, dolovanie znalostí z textov.

1 Úvod

Hodnotenie vedeckého výskumu sa v posledných rokoch stalo veľmi dôležitou činnosťou, nakoľko rozpočty organizácií zaisťujúcich financovanie vedy sa zmenšujú, ale potreba výskumu a inovácií naopak rastie. Je preto jasné, že je nevyhnutné rozpoznať vysoko kvalitný výskum, ktorý bude mať vo financovaní prioritu, od nekvalitného výskumu, ktorého podpora je neefektívna. Vedecká disciplína zaoberajúca sa meraním vedy sa nazýva scientometria a spolu so spriaznenými odbormi bibliometrie a webometrie tvorí základ prudko sa rozvíjajúceho vedného odboru zvaného informetria. Informetria stojí na rozhraní medzi informatikou a informačnou vedou a je v súčasnosti medzi vedcami veľmi aktuálnou témou [1]. Toto tvrdenie je možné dokladovať aj významom nedávno založeného Journal of Informetrics (v roku 2007), jedného z popredných časopisov v odbore informačných vied.

2 Prehľad súčasného stavu

Hodnotenie vedy je možné na rôznych úrovniach a môže byť ľahko prenesené do hodnotenia jednotlivých bádateľov, výskumných tímov, inštitúcií alebo dokonca krajín. Takéto hodnotenie sa väčšinou zakladá na hodnotení produktivity (počtu publikácií) a vplyvu (počtu citácií) výskumnej práce. V hodnotení produktivity nie sú dôležité iba počty samotných publikácií, ale aj reputácia zdrojov týchto publikácií. To nás vedie

k posudzovaniu vplyvu časopisov a konferencií. V tomto kontexte je dôležitým scientometrickým ukazovateľom kvality časopisov ich faktor vplyvu (impact factor). Ten používa len jednoduché relatívne počítanie citácií a má mnoho nedostatkov, ktoré sa informetrici snažia odstrániť.

Bollen et al. [2] aplikovali rekurzívny algoritmus PageRank používaný vo vyhľadávači Google [3] na citačnú sieť časopisov a našli veľké rozdiely medzi rebríčkami časopisov podľa kvality zisťovanej týmto spôsobom a založenej na štandardnom faktore vplyvu. Algoritmus PageRank, ktorý je možné použiť na akýkoľvek orientovaný graf, berie do úvahy nielen počet citácií získaných nejakým uzlom, ale aj kvalitu citujúcich uzlov. Kvalitný citujúci uzol má sám mnoho citácií od iných kvalitných uzlov. Preto je kvalita uzlov definovaná rekurzívne a často sa označuje za prestíž na rozdiel od popularity reprezentovanej jednoduchými počtami citácií. V porovnaní s populárnym časopisom (alebo vedcom, inštitúciou či krajinou) môže byť prestížny časopis citovaný menej, ale zato prestížnymi časopismi (vedcami). Hoci sa tieto metódy vyšších radov už dlho používajú na webe k zisťovaniu významnosti stránok, v hodnotení výskumu sú stále ešte relatívnou novinkou.

Použitie PageRanku bolo nedávno rozšírené z citačnej siete časopisov tiež na siete iných typov – bol použitý k vyhľadávaniu vynikajúcich publikácií vo fyzike a k všeobecnému hodnoteniu publikácií a krajín. PageRank a vážený PageRank boli počítané pre autorov v kocitačných sieťach, citačných grafoch a grafoch spolupráce. Vážené citácie a časový faktor boli zahrnuté v ďalších štúdiách. Vo všeobecnosti sa dá povedať, že sa PageRank ukazuje byť sľubným nástrojom hodnotenia vedeckého výstupu. Fiala a kol. [6] sa vo svojej práci zamerali na pozmenený štandardný algoritmus PageRanku, zohľadňujúci informácie nielen o citáciách medzi autormi, ale aj o ich spolupráci. Hlavnou myšlienkou je to, že nie všetky citácie majú rovnakú váhu – citácia od kolegu by mala byť považovaná za menej významnú ako citácia od cudzieho vedca. Neskôr tento model rozšírili tiež o časovú informáciu o citáciách a spolupráci [4]. V tomto novom modeli iba spolupráca predchádzajúca citácii znižuje jej váhu, zatiaľ čo počet spoločných publikácií citujúceho a citovaného autora napísaných po citácii nemá vôbec žiadny vplyv na hodnotenie citácie. Avšak počet spoločných publikácií nebol jediným faktorom ovplyvňujúcim váhy citácií – zaviedli celkom 14 nových scientometrických ukazovateľov a otestovali ich rozsiahlou kolekciou citačných dát [5].

3 Ciele projektu

Cieľom projektu je preto analýza súčasných kvantitatívnych metód hodnotenia vedeckého výskumu a návrh a overenie nových prístupov k objektívnejšiemu a spravodlivejšiemu posudzovaniu vedeckej výkonnosti. Zvláštny dôraz kladieme na metódy analýzy sietí (vrátane PageRanku a jeho variantov), v ktorých sa znalosti plzenskej textminingovej skupiny dajú výhodne skombinovať so znalosťami košickej výskumnej skupiny, ktorej členovia v minulosti prevádzali analýzy okrem iného firemných a citačných sietí [7], [8].

Cieľom projektu je tiež riešenie problémov ako napr.:

- rozlíšenie medzi celoživotnými zásluhami a súčasnou výkonnosťou,
- zohľadnenie spoluautorstva ako v publikáciách tak v citáciách,
- zohľadnenie rozdielov medzi jednotlivými vedeckými odbormi,
- odlišné správanie sa vedcov v rôznych fázach ich kariéry a ďalšie.

4 Dosiahnuté výsledky

V článku [9] sme skúmali možnosť využívania prepojených dát za účelom pokročilej analýzy softvérových špecifikácií. Tieto dokumenty sa svojou odbornosťou a použitým názvoslovím veľmi podobajú vedeckým publikáciám. S úspechom je teda možné ich využívať pre vývoj metód, ktoré budú následne aplikované na vedecké články. Môže sa jednať napr. o detekciu pomenovaných entít, ale hlavne o odvodzovanie témy článku podľa nájdených pojmov a určenia vzdialenosti medzi článkami v priestore prepojených dát. Scientometriu je tak možné obohatiť o automaticky určené tematické oblasti článkov a autorov je možné automaticky deliť podľa ich oblasti záujmu, bez toho aby sme boli závislí na správnej voľbe kľúčových slov a kategórií pri vedeckých publikáciách.

Ďalším naším výsledkom je článok [10], v ktorom skúmame otázku, či je vhodné hodnotiť autorov podľa siete autorov alebo siete publikácií. Za týmto účelom využívame niekoľko variant PageRanku a vyhodnocujeme ich s využitím dát z ISI Web of Science.

V inej práci [11] sme sa zaoberali vzťahom medzi PageRankom a jednoduchým počítaním citácií ako vhodných ukazovateľov významnosti vedcov a v ďalšej práci sme sa venovali vplyvu starnutia hrán v sieti [12], t.j. redukcii zriedkavých a naopak zosilňovania častých a významných hrán v citačných a kolaboračných sieťach autorov na hodnotenie úspešnosti výskumníkov [13]. Zo všetkých uvedených štúdií bolo najväčšie množstvo dát spracovaných v [11], kde sa spracovával citačný graf s viac ako pol miliónom publikácií niekoľkými miliónmi citácií medzi autormi. Aj tak sa ale analýza dala realizovať bežnými výpočtovými prostriedkami.

Vyhodnocovanie efektivity skúmaných metód oceňovania kvality vedeckých pracovníkov je vo všetkých prípadoch pomerne chúlostivou záležitosťou a spočíva v automatizovanom vytváraní rebríčkov autorov odborných publikácií na základe uvedených informatických metód a v ich porovnávaní s určitým referenčným rebríčkom – zlatým štandardom úspešných vedcov. V našich experimentoch sme za tento zlatý štandard považovali množinu vedcov, ktorí dostali nejaké prestížne ocenenie (napr. ACM Turing Award) alebo pôsobia v edičných radách významných časopisov vo svojom odbore. Výsledky vyššie uvedených troch publikácií [10, 11, 13] je možné zhrnúť konštatovaním, že PageRank všeobecne (vzhľadom k svojim výpočtovým nákladom) nemusí dávať lepšie výsledky než jednoduché počítanie citácií, že je vhodnejšie ho počítať zo siete publikácií než zo siete autorov a že vplyv starnutia hrán v kolaboračnej sieti autorov sa v niektorých prípadoch prejavuje pozitívne objektívnejším ohodnotením významu autorov.

Okrajovo sme sa venovali aj možnostiam vhodnej vizualizácie výsledkov našich algoritmov v rámci danej citačnej siete [14].

5 Poďakovanie

Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe Zmluvy č. SK-CZ-2013-0062 a grantom MSMT MOBILITY 7AMB14SK090.

Literatúra

1. Bar-Ilan, J.: Informetrics at the beginning of the 21st century-A review. *Journal of Informetrics*, 2 (2008), 1-52.
2. Bollen, J., Rodriguez, M. A., Van De Sompel, H.: Journal status. *Scientometrics*, 69 (2006), 669–687.

3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30 (1998), 107-117.
4. Fiala, D.: Mining citation information from CiteSeer data. *Scientometrics*, 86 (2011), 553-562.
5. Fiala, D.: Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, 6 (2012), 370-388.
6. Fiala, D., Rousselot, F., Ježek, K.: PageRank for bibliographic networks. *Scientometrics*, 76 (2008), 135-158.
7. Repka, M., Paralič, J.: *Company Networks Analysis*. LAP Lambert Academic Publishing, Saarbrucken, 2013.
8. Repka, M., Paralič, J.: Objavovanie znalostí v citačných sieťach. In *Proc. ZNALOSTI*, Pavel Smrž (Ed.), VŠE v Praze, Nakladatelství Oeconomica (2010), 247-250.
9. Dostal, M., Nykl, M., Ježek, K.: Semantic analysis of software specifications with Linked Data in *Journal of Theoretical and Applied Information Technology*, 67 (2014), 368-376.
10. Nykl, M., Ježek, K., Fiala, D., Dostal, M.: PageRank variants in the evaluation of citation networks. *Journal of Informetrics*, 8 (2014), 683-692.
11. Fiala, D., Šubelj, L., Žitnik, S., Bajec, M.: Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, 9 (2015), 334-348.
12. Tutoky, G., Paralič, J.: Time Based Modelling of Collaboration Social Networks. *Lecture Notes in Computer Science*, 6922 (2011), 409-418.
13. Fiala, D., Tutoky, G., Koncz, P., Paralič, J.: Ageing of edges in collaboration networks and its effect on author rankings. *Acta Polytechica Hungarica* (submitted in 2015).
14. Kováčová, T., Havrilová, C., Paralič, J.: Návrh a implementácia vizualizácie citačných sietí. *Electrical Engineering and Informatics VI* (submitted in 2015)

Annotation:

Modern informetric methods for the evaluation of scientific research

This paper briefly presents Czech-Slovak research project focussed on analysis of present, as well as the design and verification of new scientometric indicators based on citation network analysis methods and text mining methods. Within citation networks analysis methods the main focus is on adaptation options of PageRank algorithm for the needs of given goal. Within methods based on text mining the focus is on modelling the relationship between scientometric indicators of publications significance and their attributes obtained by text mining methods. The object of the project is also exploring dependencies between citation rates and the popularity of the topic, as well as visualization of citation networks.