Analýza kvality konferencií pomocou komunít výskumníkov

Pre digitálne knižnice v odbore informačných a komunikačných technológií (ďalej len IKT) konferencie zohrávajú významnú rolu v šírení publikačných a vedeckých výstupov. Avšak pri aktuálnom výraznom rozvoji konferencií je čoraz náročnejšie z pohľadu výskumníkov a knihovníkov posudzovanie kvality týchto konferencií. V našej práci sa venujeme metódam automatizovaného odhaľovania prestížnych konferencií analýzou rôznych online bibliografických databáz, ktoré obsahujú veľké množstvo informácií o výskumných aktivitách z rôznych oblastí. Dané online databázy vytvárajú významné informačné siete, spájajú výskumné práce, autorov, konferencie/časopisy, ako aj citačné informácie. V našej práci sa zaoberáme novými prístupmi klasifikácie konferencií, a to pomocou komunity výskumníkov získaných z bibliografických informácií. Využívame pritom existujúce algoritmy na odhaľovanie komunít pre samotné hodnotenie kvality konferencií.

Úvod

Samotný potenciál digitálnych knižníc (ďalej len DK) nespočíva len v sprístupňovaní obrovskej zbierky vedomostí, ale tiež v poskytovaní efektívnych nástrojov pre používateľov na odporúčanie a filtráciu obsahu. Keď používateľ či knihovník prehľadáva literatúru alebo vykonáva rôzne rozhodnutia, využívajú sa práve bibliometrické údaje na určenie kvality dokumentov. DK v oblasti IKT, akými sú napr. ACM Portal a CiteSeer, potrebujú merať kvalitu vedeckých konferencií automatizovaným spôsobom. Táto úloha je netriviálna z dvoch dôvodov. Po prvé IKT je unikátna disciplína, čo sa týka publikačnej činnosti. Na rozdiel od ostatných oblastí konferencie hrajú rovnako dôležitú rolu (ak nie vyššiu) ako karentované časopisy. Je to z toho dôvodu, že samotná disciplína má veľmi rýchlo sa rozvíjajúci charakter, čo si vyžaduje progresívne rozšírenie nových vedeckých poznatkov. Pri miere prijatia príspevkov 10 – 20 % publikácie z konferencií často získavajú viac citácií ako časopisy [1], navyše prestížne konferencie často patria medzi najdôležitejšie miesta prezentovania vedeckého výstupu.

Po druhé s rýchlym rozvojom IKT súvisí aj rapídny rozvoj konferencií v posledných rokoch, čo je preukázateľné aj v našich získaných údajov z DBWorld[1]. Pre výskumníkov a knihovníkov sa stáva čoraz viac dôležitejšou práve reputácia (teda kvalita) konferencií. Problém však spočíva v tom, ako automaticky rozoznať kvalitu medzi stovkami zverejnených *„Call for Papers*" (CFP) každoročne.

V IKT množstvo dostupných dát na webe umožňuje vytvárať mechanizmy na automatizované hodnotenie kvality. Využívať jediné meradlo na hodnotenie kvality, akým je napr. *h-index* [2], môže byť pomerne nebezpečné, hlavne ak sa meradlo použije bez zváženia. H-index je jednoduché meradlo na podporu hodnotenia s cieľom odstrániť zložité aspekty rozhodovacieho procesu, ale použitie iba jedného číselného ukazovateľa kvality nie je akceptovateľnou cestou pre hodnotenie výstupu výskumníka.

Súčasné techniky merania reputácie a kvality miest na publikovanie využívajú rôzne citačné metriky, ako je napr. *impakt faktor* (IF) [3]. Podľa našej mienky však uvedené jednoduché metriky nie sú dostačujúce pre meranie kvality IKT konferencií z nasledujúcich dôvodov:

- Historické citačné štatistiky nie sú k dispozícii pre novo vznikajúce alebo mladé konferencie, čím metriky na báze citácií sú nepoužiteľné.
- Aj pre už tradičné konferencie si citačné štatistiky vyžadujú určitý čas na akumuláciu. Štúdie z hlavných databáz konferencií a časopisov ukazujú, že väčšina citácií pribúda v rozpätí päť a viac rokov [1].
- Keď výskumník skúma CFP alebo prechádza webové stránky konferencie a samotná udalosť nie je v danom odbore dobre známa, je pravdepodobné, že samotný výskumník bude mať výrazné problémy pri získaní citačných štatistík danej konferencie, teda aj pri určení kvality konferencie.

V danej práci sa venujeme rôznym heuristikám na nájdenie korelácie medzi charakteristikami publikujúcich a kvalitou konferencií. Vzhľadom na veľkú zbierku CFP z konferencií využívame rôzne techniky extrakcie na rozpoznanie a extrahovanie mien autorov.

1 Súčasný stav výskumu v oblasti IKT

Meranie kvality publikačných miest je dôležitou úlohou v bibliometrii. Najviac rozšíreným spôsobom merania kvality je Garfieldov *ukazovateľ priemernej citovanosti – impakt faktor* (IF) [3]. Ide o číselné vyjadrenie podielu medzi celkovým počtom citácií článkov daného časopisu alebo konferencie za dva roky a celkovým počtom článkov publikovaných v tomto časopise alebo konferencie za rovnaké obdobie. IF od samotného zavedenia je často kritizovaný predovšetkým kvôli jeho závislosti od počtu citácií [4], a preto vzniklo mnoho alternatív, napr. *h-index* [2], *rôzne metriky na báze PageRank* [5] a *metriky na báze sťahovaní* [6] na hodnotenie vedeckých časopisov v odbore IKT [7].

Pre hodnotenie výskumníka ako jednotlivca spomenutý h-index bol navrhnutý J. E. Hirschom [2]. Samotné meradlo je pomerne zaujímavé, má však určité nevýhody. Podľa Aparecida~\cite{aparecida:oaqa} najväčším problémom je nejasnosť spôsobená

[1] http://dbis-group.uni-muenster.de/dbms/templates/conferences/conferences.php

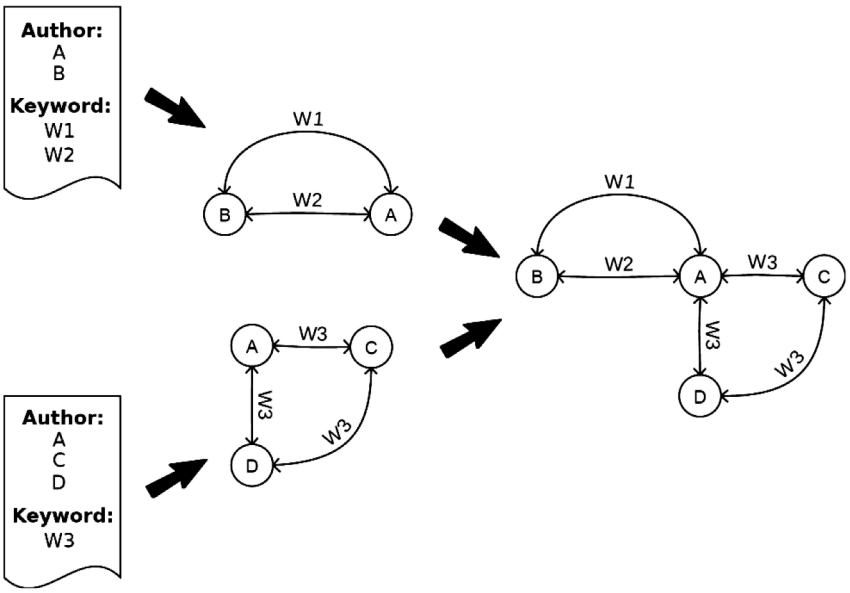
priradením vedeckej hodnoty ku popularite, keďže povrchný článok môže byť populárnejší ako iný so skutočným vedeckým prínosom. Ďalej môže byť impakt index postihnutý nevýhodou pozitívnej spätnej väzby. Je pomerne známy a častý jav, že používatelia prezerajú iba prvú stránku nájdených výsledkov pri vyhľadávaní publikácií, teda najviac citovaní autori sa stanú ešte viac citovanejšími (napr. články objavujúce sa na prvej strane Google Scholar).

Ďalší nedávno objavený zaujímavý jav [11] hovorí o tom, že existuje korelácia medzi počtom autorov publikácie a jeho citovanosťou. Čím väčšie skupiny autorov publikujú rôzne články, tým viac citácií daný článok získava. Existuje tiež silná tendencia citovania publikácií potenciálnymi recenzentmi alebo kolegami z rovnakej krajiny alebo pracoviska. Autori preukázali v odbore strojárenstva zvýšenie počtu citácií publikácií s viac ako 5 autormi až o 3,72-krát, v porovnaní s publikáciami od jedného autora táto hodnota sa zvýši až na 13,01 násobok. Napriek uvedeným nevýhodám využívanie metriky h-index nám poskytne ľahko vyčísliteľnú hodnotu na odhadnutie významnosti vedeckého výstupu.

2 Siete výskumníkov

Na riešenie problémov známych hodnotiacich metód navrhujeme nové metódy na základe analýzy komunít výskumníkov zúčastnených na konferenciách, a teda kvalifikovanie konferencií na základe ich hlasovania. Výskumná (vedecká) komunita môže byť definovaná ako rôznorodá sieť vzájomne prepojených vedcov pracujúcich v konkrétnych vedeckých odboroch a v určitých inštitúciách. Členovia komunít sa môžu zúčastniť na konferenciách ako prezentujúci alebo ako člen programového výboru. Aktuálne sa zameriavame na prvú skupinu výskumníkov.

Na nájdenie výskumnej komunity používame vzťah spoluautorstva publikácií, pretože tento vzťah je základným staveným kameňom výskumnej skupiny, navrhovaný v [9]. Začnime jednoduchým modelom publikácie. Model publikácie sa skladá z kľúčových slov a mien autorov. V tomto prípade môžeme vytvoriť výskumný záujem autorov z kľúčových slov v autorových dielach. Autori, ktorí napíšu článok spoločne, zdieľajú rovnaký záujem, pričom tento záujem prezentujú práve kľúčové slová. Túto bibliografickú informáciu môžeme reprezentovať ako sieť výskumníkov.



Obrázok č. 1: Model siete výskumníkov

Vysvetlíme si túto metódu na príklade. Predpokladajme, že máme dva články, ako je reprezentované na ľavej strane obrázka č. 1. Daný článok bol napísaný dvoma autormi A a B a má dve kľúčové slová – W_1 a W_2. Druhý bol napísaný troma autormi A, B a C a obsahuje kľúčové slovo W_3. Môžeme vytvoriť grafy autorov, ako je uvedené v strede obrázka č. 1. Následne sa vytvorí celý graf, ktorý je spojením grafov dvoch uvedených dokumentov. Komunitu výskumníkov teda definujeme ako zoskupenie, ktoré je husto prepojené pomocou rovnakého výskumného záujmu. Pomocou tejto metódy môžeme vytvoriť dátovú sadu komunít výskumníkov. Vytvorený graf sa môže výrazne zväčšiť, ak sa analyzuje veľké množstvo autorov a článkov.

3 Kolekcia dát

Pri navrhovaní a overovaní danej metódy sme využili rôzne sady dát. Dáta o výskumníkoch sme získali z databázy DBLP[2] pomocou metódy uvedenej v [10].

Na základe rebríčka konferencií hodnotených pomocou IF z portálu ConferenceRanking.org[3] sme extrahovali vysoko hodnotené konferencie v každom odbore IKT a získali ich celé názvy z DBLP. Následne sme extrahovali z dátovej sady DBWorld najnovšie CFP, ktoré približne zodpovedali názvom týchto špičkových konferencií. Výsledné CFP tvorili reprezentatívnu množinu na porovnanie s hodnotenými konferenciami.

4 Hodnotenie konferencií

Pri dolovaní bibliografických databáz na efektívne hodnotenie záznamov sme použili nasledujúce heuristiky. Heuristiky slúžia na nastavenie váhy konferencií/autorov, časopisov a oblastí výskumu. Tieto váhy sa nastavujú v opakovanom procese, ktorý končí vtedy, ak sa váhy ustália.

- *Konferencia/časopis má dobrú reputáciu, ak ovplyvní veľa dokumentov od vysoko hodnotených autorov*

1. Autor sa považuje sa vysoko hodnoteného, ak publikuje mnoho článkov v renomovaných časopisoch a na konferenciách.

[2] http://www.informatik.uni-trier.de/~ley/db/
[3] http://www.conference-ranking.org/

2. Autori často publikujú na rovnakých konferenciách, ak tie zdieľajú rovnaké alebo podobné záujmy výskumu.
3. Konferencia/časopis patrí do jednej oblasti výskumu, ak získala dokumenty od vysoko hodnotených autorov pôsobiacich predovšetkým v tejto oblasti.
4. Skupina konferencií/časopisov patrí do rovnakej oblasti, pokiaľ zverejňujú dokumenty najmä tejto oblasti.

Aby sme mohli hodnotiť konferencie pomocou komunít, musíme jednotlivých členov komunity kategorizovať. Na dosiahnutie toho sme extrahovali vysoko hodnotených jednotlivcov zo skupín. Algoritmus na určenie hodnotenia jedinca berie do úvahy publikácie jedinca, počet citácií danej osoby a dobu, počas ktorej publikuje v danom odbore. Extrahovali sme pomocou tohto prístupu vzorku publikácií, autorov a spoluautorov. Dáta boli následne uložené do relačnej databázy.

Vysoko hodnotení autori sa určujú na základe váh, ktoré sa k nim priradia. Využívame tri váhy: váha publikácií, váha citácií a váha recenzentov, ktoré sa rátajú nasledovným spôsobom:

Váha publikácií = Počet publikácií / Obdobie (počet rokov)
Váha citácií = Počet prijatých citácií od 1 autora / Celkový počet citácií
Váha recenzentov = Počet recenzentov / Celkový počet autorov
Celková váha = váha publikácií + váha citácií + váha recenzentov

Vysoko hodnotení autori sú teda určení na základe ich celkovej váhy. Dolovaním účastníkov konferencií a hľadaním vysoko hodnotených autorov medzi nimi dokážeme určiť počet expertov zúčastnených sa na konferenciách. Čím vyššia váha sa priraďuje účastníkom konferencie, tým vyššie hodnotenie získa samotná konferencia.

Generovanie dát o komunitách, rátanie ich expertných váh a optimalizácie atribútov algoritmu aktuálne prebieha, podarilo sa nám však dosiahnuť tieto výsledky:

- získali a spracovali sme rôzne dáta na identifikovanie výskumnej komunity,
- v získaných dátach sme identifikovali výskumné komunity využitím ich publikácií,
- definovali sme rôzne heuristiky hodnotenia konferencií.

5 Zhrnutie

V tomto článku sme sa zaoberali rôznymi metódami na identifikovanie povesti konferencií pomocou dolovania charakteristík výskumných komunít. Pri kombinácii s klasifikačnými schémami by tieto heuristiky mohli dosiahnuť zaujímavé výsledky a dostatočnú presnosť pri rozlišovaní kvality konferencií.

Aktuálne analyzujeme iba ukončené konferencie, ale dolovanie komunít by mohlo byť užitočné aj pri predvídaní kvality vznikajúcich konferencií analýzou členov programového výboru, keďže iné informácie nie sú k dispozícii pre nové konferencie. Naša metóda má určité obmedzenia, čo sa týka zložitosti vytvorenej dátovej sady, čo môže viesť k neustálej zmene váh v každej iterácii výpočtu.

Veríme, že náš výskum môže byť užitočný pre výskum bibliometrie zavedením nových prístupov k meraniu kvality publikačnej činnosti. V súčasnosti ešte musíme splniť niekoľko kritérií a vyriešiť rôzne problémy na generovanie relevantných výsledkov, porovnateľných s existujúcimi klasifikačnými metódami.

Použitá literatúra

[1] RAHM, E. a A. THOR. Citation analysis of database publications. SIGMOD Record, Vol. 34, No. 4 (2005).
[2] HIRSCH, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, Vol. 102(46), 16569-16572 (2005).
[3] GARFIELD, E.: Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, Vol:122, No:3159, p. 108-111 (1955).
[4] SAHA, S., SAINT, S. a D. A. CHRISTAKIS. Impact factor: a valid measure of journal quality? In: *Journal of the Medical Library Association*, Vol. 91(1): 42-46 (2003).
[5] BOLLEN, J., RODRIGUEZ, M. A. a H. Van de SOMPEL. Journal Status. Dostupné na: http://www.arxiv.org/abs/cs.GL/0601030 (2006)
[6] BOLLEN, J., Van de SOMPEL, H., SMITH, J. a R. LUCE. Toward alternative metrics of journal impact: A comparison of download and citation data. Information Processing and Management, 41(6): 1419-1440 (2005)
[7] NERUR, S., SIKORA, R., MANGALARAJ, G. a V. BALIJEPALLY. Assessing the Relative Influence of Journals in a Citation Network, Communications of the ACM, Vol.48, No.11, (2005).
[8] APARECIDA, M., WARPECHOWSKI, M. a J. PALAZZO. An ontological approach for the quality assessment of computer science conferences. Proceedings of the 2007 conference on Advances in conceptual modeling: foundations and applications, ER'07. Springer-Verlag: 202-212 (2007).
[9] NEWMAN, M.E.J.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the USA, 101(suppl. 1), 5200-5205 (2004).
[10] LEY, M.: DBLP: some lessons learned. Proceedings of the VLDB Endowment, VLDB Vol. 2 issue 2., 1493-1500 (2009).
[11] WUCHTY, S., JONES, B.F. a B. UZZI. The Increasing Dominance of Teams in Production of Knowledge. Science (2007).

Zoltán Harsányi
harsanyi@fiit.stuba.sk

Viera Rozinajová
rozinajova@fiit.stuba.sk