Objavovanie expertných komunít

Peter Hamar*, Lukáš Sekerák#

*# Fakulta informatiky a informačných technológií,
Slovenská technická univerzita v Bratislave, Ilkovičova 2, 842 16 Bratislava 4

*Petko.hamar@gmail.com , #sekys.sekys@gmail.com

Abstract. Na internete sa nachádza veľké množstvo vedeckých databáz. V týchto databázach sú priamo obsiahnuté dáta o autoroch, ich prácach a o dielach na ktoré sa v nich odvolávali. Keď sa dobre zamyslíme, je možné z týchto dát pomocou data mining-u získať nové informácie. Je možné sa zamerať na rôzne oblasti, ako je hodnotenie autorov, publikácií, konferencií, hľadanie expertov, profilovanie atď. Jednou pre nás veľmi zaujímavou časťou je objavovanie komunít. Ako dátový set sme si zvolili podmnožinu databázy DBLP, na ktorej sme sa pokúsili aplikovať experimentálny algoritmus Fuzzy K-Means. Rovnakú množinu sme pretransformovali do grafu a vizualizovali sme ho pomocou nástroja Gephi. Výsledky sme porovnali.

Keywords. Zhlukovanie, vedecké komunity, DBLP, data mining.

1 Opis problému

Na internete sú dostupné najrôznejšie databázy ktoré zoskupujú vedecké práce. Známymi sú hlavne databázy ACM, Springer, IEEE, DBLP ale aj iné. Všetky takéto databázy obsahujú veľké množstvo informácií vhodných na skúmanie. Pod skúmaním myslíme použitie data-mining-u v podobe upraveného K-Means zhlukovania.

Prakticky každý autor spolupracuje na písaní s inými autormi. Zdieľa s nimi myšlienky, názory, spoločne sa podieľajú na písaní diel a tvoria tak akúsi komunitu. Samotné komunity nie sú priamo v týchto databázach obsiahnuté, avšak pomocou spoluautorstva ich je možné získať. Veľmi žiadúcim pri objavovaní týchto komunít je vziať do úvahy aj dimenziu času. Komunity sa v čase menia a vyvíjajú. Ľudia ktorý spolupracovali pred 20-imi rokmi už v súčasnosti spolupracovať nemusia.

Samotné definovanie komunít však nie je triviálne. Existujú rôzne spôsoby a metódy objavovania komunít. Každá metóda je niečím špecifická. Pokiaľ by bolo možné toto zhlukovanie vykonať v dostatočnej kvalite, následne by sa to dalo využiť napr. v rámci vyhľadávania expertov vo zvolenej problémovej oblasti. Táto práca sa venuje problematike hľadania expertných komunít, zahŕňa výber metódy na objavovanie komunít. Cieľom je implementovanie a otestovanie tejto metódy, ktorá bude zohľadňovať typ práce, množstvo autorov pri písaní prác a taktiež vezme v úvahu vyvíjanie komunít v čase.

2 Opis prác iných autorov

V práci „A new K-means algorithm for community structures detection based on Fuzzy clustering“ tvrdia, že jedným z hlavných problémov pri určovaní komunít je to, že jeden autor môže patriť do viacerých komunít a je náročné určiť vzdialenosti od stredov týchto komunít. Rozhodli sa využiť tzv. Fuzzy K-Means. Avšak tento spôsob je náchylný na správne určenie počiatočnej hodnoty, ak táto hodnota nie je zvolená správne tak konverguje k lokálnemu minimu. Zvyšok práce zahŕňa detailné vysvetlenie algoritmov a ich otestovanie. [2]

Pri objavovaní komunít sa taktiež používa tzv. Newmanov algoritmus. Tento algoritmus bol použitý v práci[4]. Taktiež ho je možné použiť aj na zhlukovanie dokumentov. Definovali si 3 entity: autora, dokument a termín – kľúčové slovo. V ďalšej časti ich práce detailnejšie popisujú princíp objavovania prepojení a samotného zhlukovania pomocou Newmanovho algoritmu.

3 Predspracovanie a výber atribútov

Pre náš problém sme si zvolili dátový set DBLP , ide o jednu z najväčších vedeckých databázach. Tento dátový set je kompletne uložený v XML súbore. Veľkosť tohto súboru je viac ako 1.3 Gb a zoskupuje viac ako 1,3 milióna publikácií. Na začiatku práce bolo potrebné zanalyzovať dáta v súbore. Zistiť či sú postačujúce pre riešenie nášho problému. Štruktúra tohto XML súboru je definovaná pomocou DTD(Document type Definition) súboru , ktorý detailne opisuje dáta.

Tento dátový set zoskupuje najrôznejšie druhy prác, konkrétne: Article, inproceedings, proceedings, book, incollection, phdthesis, mastersthesis, www. Všetky tieto typy majú definované spoločné atribúty a v prípade, že nejakej práci chýba jeden z atribútov, tak nie je uvedený. Príklady atribútov: author, editor, address, title, booktitle, pages, year, journal, volume, number, month a mnoho ďalších. Dátové typy jednotlivých atribútov sú definované v priradenom DTD súbore, konkrétne ide o typ #PCDATA, teda o text. Každý element v XML súbore, je jednoznačne definovaný kľúčom „key“.

Mnohé tieto údaje nebudeme potrebovať, respektíve závisí to aj od spôsobu určenia úlohy. Pre našu úlohu sme zvolili atribúty:

- Typ publikácie
- Dátum vydania publikácie
- Množina autorov publikácie

Proces výberu vlastností objektov a ich mapovania na čísla je známy ako vektorizácia[3]. Tomuto procesu sa venujeme v ďalšej časti.

3.1 Ukážka dát

Inproceedings:

<inproceedingsmdate="2012-03-09" key="conf/globecom/RathRB11">
<author>HemantKumarRath</author>
<author>M. A. Rajan</author>
<author>P. Balamuralidhar</author>
<title>MonotonicSignedGraphapproachforcross-layercongestioncontrol in wireless ad-hoc networks.</title>
<pages>309-314</pages>
<year>2011</year>
<booktitle>GLOBECOM Workshops</booktitle>
<ee>http://dx.doi.org/10.1109/GLOCOMW.2011.6162459</ee>
<crossref>conf/globecom/2011w</crossref>
<url>db/conf/globecom/globecom2011w.html#RathRB11</url>
</inproceedings>

3.2 Predspracovanie dát

Na prvotnú analýzu sme použili nástroj ktorý dokáže otvoriť veľký xml súbor (010 editor). Následne sme si vytvorili vlastné histogrami pre typy publikácií, pre roky vydania publikácií a pre počet autorov na publikáciu.

Na základe histogramov sme spoznali skutočné typy údajov a ich rozsah. Z histogramu(obr. 1) môžeme vidieť, že najviac prác je takých, kde počet spoluautorov je 0 tzv. na vytváraní publikácie sa podieľal práve 1 autor. Preto publikácie ktoré nemajú autora, alebo majú len jedného nebudeme v práci brať v úvahu (pretože publikácia nemá spoluautorov).

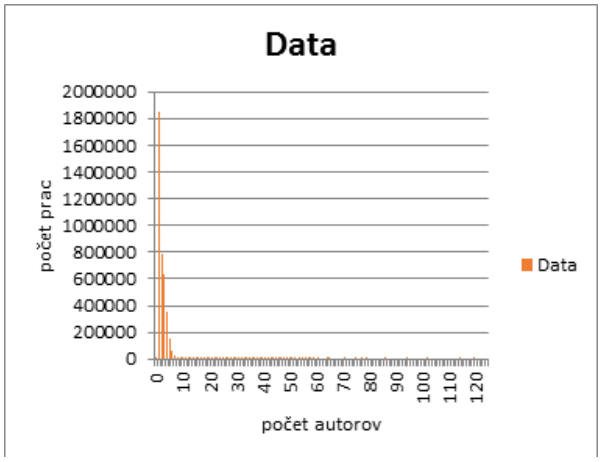

Fig. 1. Histogram počtu spoluautorov.

Keďže našou úlohou je hľadanie komunít, tak chápeme, že každá inštancia(autor) má N dimenzií. Kde každá dimenzia reprezentuje spojenie s iným autorom. Výsledkom je teda matica o veľkosti NxN kde N je počet autorov. Naše vybrané atribúty priamo ovplyvňujú silu spojenia dvoch autorov. Teda výsledná váha atribútov je uložená v bunke matice.

Na vytvorenie tejto matice sme použili vlastný program, ktorý oparsoval vstupné xml. Vybral zvolené atribúty pre publikáciu, vypočítal ich celkovú váhu a priradil ju ku každému autorovi publikácie.

V rámci procesu vektorizácie musíme najprv priradiť jednotlivým druhom publikácií určitú váhu. Určenie váh publikácií je zhrnuté v tabuľke.

	book	phdthesis	mastersthesis	proceedings
Váha diela	1	0.9	0.8	0.6
	Inproceedings	Incollection	article	www
Váha diela	0.4	0.4	0.4	0.2

Faktor času zahrnieme do hodnotenia tak, že na základe dátumu publikácie si určíme ako dávno bola práca publikovaná. Získanú hodnotu priradíme do určitého intervalu a následne mu podľa daného intervalu určíme hodnotu.

	0 – 2 roky	3- 5	6 - 10	10 -
Váha	1	0.75	0.5	0.25

Následne bude hodnotenie predelené počtom autorov, lebo v prípade ak je autorov viac, tak je potrebná menšia miera spolupráce. Hodnotenie sily spojenia 2 autorov pre 1 publikáciu je:

$$x = v_d * v_c / p_a$$

v_d – váha typu publikácie, p_a – počet autorov, v_c – váha času/novosti

Počet autorov sa v každej publikácii líši a nechceme aby jedna dimenzia mala väčší vplyv ako iná. Preto tento atribút normalizujeme. Finálne hodnotenie sily spojenia dvoch autorov na pozícii i, j je:

$$x_{i,j} = \sum_{0}^{k} x_{i,j}$$

Keďže proces tvorby matice je výpočtovo náročný a pracujeme s obrovským množstvom dát rozhodli sme sa použiť iba určitú podmnožinu dát. Aj napriek tomu, že používame iba podmnožinu museli sme starostlivo vyberať také algoritmy na spracovanie, ktoré majú časovú zložitosť O(logn).

4 DM metódy

Fuzzy K-means je nová experimentálna metóda na objavovanie komunít. Metóda je založená na pôvodnom K-means, pričom využíva fuzzy logiku. To znamená, že každá inštancia môže patriť do viacerých zhlukov, s určitou pravdepodobnosťou. Rovnako ako K-means využíva parametre: vstupné dáta, vstupné klastre, počet zhlukov, maximálny počet iterácií, deltu konvergencie a fuzzy faktor[1]. Viac o tomto algoritme je popísané v článku[2]. Tento algoritmus tiež vyžaduje vstupné klastre, ktoré sú definované ich pozíciou. Existujú rôzne rozšírenia, ktoré túto pozíciu vypočítajú. Naše riešenie vo frameworku Mahout túto pozíciu zvolilo náhodne. Preto je tento algoritmus citlivý na výber počiatočnej hodnoty pozície. Z toho dôvodu je možné, že pri opakovanom spustení dostaneme rôzne výsledky.

Pri testovaní tejto metódy, sme prakticky vždy dostali neprijateľné výsledky. Pretože súčet pravdepodobností bol väčší ako 1. Po analýze dokumentácií sme zistili, že ide o chybu v implementácií knižnice Mahout. Tuto metódu sme preto zavrhli.

Následne sme našu úlohu charakterizovali ako druh grafového problému. Takže sme k naším spracovaným dátam pristupovali ako ku vrcholom (autori) a hranám. Na prácu s grafom sme použili nástroj Gephi, ktorý dokáže v grafe hľadať komunity. Tento nástroj opisujeme v samostatnej kapitole. Rovnako aj celý postup objavovania komunít v grafe. Našu ohodnocovaciu funkciu sme naďalej použili na ohodnotenie váh hrán.

Zároveň sme sa pokúsili použiť algoritmus K-Means namiesto pôvodného Fuzzy K-Means. Táto metóda má podobné parametre ako sú uvedené v úvode tejto kapitoly. Na použitie tejto metódy sme opäť využili Mahout. Aby sme mohli použiť našu doterajšiu implementáciu a nemuseli sme použiť ďalší novy formát dát.

5 Vizualizácia pomocou Gephi

Na vizualizáciu formou grafu sme použili nástroj Gephi. Je dostupný na stránke[2]. Tento nástroj okrem vizualizácie dokáže aj mnohé ďalšie veci, jednou z nich je aj detekcia komunít. Na detekciu komunít sa používa algoritmus Modularity[1].

Tento algoritmus skalárne rozdeľuje hodnoty medzi <-1,1> na základe hustoty spojení v rámci komunity. Skladá sa z dvoch fáz ktoré sa iteratívne opakujú. Na začiatku máme ohodnotený graf. Každému vrcholu grafu priradí iná komunita. Takže na začiatku máme toľko komunít koľko uzlov. Potom sa pre každý uzol(komunitu) i hľadajú susedia j. Počíta sa hodnota modularity, ktorá určuje mieru zisku, ktorý získame odstránením uzla i z jeho súčasnej komunity a umiestnením v rámci nájdenej komuni-

[1] http://en.wikipedia.org/wiki/Data_clustering#Fuzzy_c-means_clustering
[2] https://gephi.org/

ty j. Po prejdení všetkých, sa uzol i následne umiestni do komunity j, kde bol vypočítaný maximálny zisk, avšak iba ak je pozitívny. Ak zisk nie je pozitívny tak tento uzol ostáva vo svojej pôvodnej komunite. Tento proces sa opakuje kým sa nedosiahnu lokálne maximá modularity teda, keď už ďalšie presuny neprinesú zvýšenie zisku. Druhá fáza algoritmu spočíva vo vytvorení novej siete ktoré uzly sú teraz nové detegované komunity počas fázy 1.

Predspracované dáta sme pretransformovali do formátu ktorý používa nástroj Gephi. Na samotnú vizualizáciu bolo potrebné vykonať viaceré nastavenia filtrov, zobrazení a algoritmov. Presný postup vizualizácie pomocou tohto nástroja:

1. Dáta otvoríme v nástroji Gephi (na začiatku nie sú komunity viditeľné)
2. Ako spôsob rozloženia (layout) zvolíme Force atlas (je potrebné to zastaviť v momente keď sa pozície bodov ustália)
3. Na ohodnotenie vrcholov zvolíme možnosť „number of connections“, čo nám zaručí, že sa vrcholy ohodnotia na základe počtu spojení.
4. Následne určíme priemernú vzdialenosť vrcholov v grafe, povie nám to ako sú ďaleko vrcholy od seba vzdialené. Použijeme na to metriku „average path length“.
5. Potom zmeníme hodnotiaci parameter na „betweeness centrality“, taktiež nastavíme veľkosť uzlov.
6. Zaškrtneme voľbu „adjust by sizes“ v časti layout.
7. Zobrazeným vrcholom zviditeľníme mená, pomocou tlačidla T a nastavíme štýl písma.
8. Na detekciu komunít následne použijeme algoritmus Modularity, z okna statistics. Následne ho aplikujeme na graf.
9. Posledným krokom je skrytie menej významných komunít, pomocou filtra. Zvolíme filter „degree range“ a nastavíme rozsah.

6 Spôsob vyhodnotenia

V časti DM metódy sme prezentovali 3 metódy. Jedna z metód Fuzzy K-Means nemôže byť vyhodnotená pre abnormálne výsledky. Preto sa nebude brat v úvahu pri vyhodnotení.

Ďalej sme spomenuli metódu, hľadanie komunít v grafe za pomoci Gephi nástroja. Výsledkom takého hľadania je vizualizácia, ktorá nám má poukázať na expertné komunity. Na vizualizovaný graf však nemôžme použiť metriky a nemôžme tak zmerať výsledok hľadania. Preto túto metódu budeme považovať za náš „zlatý štandard“.

Ako tretiu metódu sme uviedli K-Means. S touto metódou budeme experimentovať a budeme hľadať najlepšie parametre, aby sa jej výsledok priblížil k nášmu „zlatému štandardu“.

Tieto 2 metódy teda porovnáme iba vizuálne. Pre výsledok K-Means metódy sa pokúsime vypočítať Davies-Bouldin Index. Skúsime ho použiť, aby sme sa naučili pracovať aj s indexami a overili naše výsledky.

7 Expertimenty

Na vizualizáciu použijeme vyššie spomínaný nástroj Gephi. Tento nástroj ponúka algoritmus Modularity[1] na detekciu komunít. Keďže naša dátová sada bola značne rozsiahla bolo potrebné túto sadu vhodným spôsobom zredukovať. Redukciu sme sa rozhodli vykonať na základe mesiaca a dňa vydania publikácie. To z dôvodu, aby sme deterministicky vyberali dáta, nie náhodne. Pomohlo by nám to vybrať rovnakú podmnožinu dát pri opakovanom spúšťaní. Redukcia na základe mesiaca a dňa nám neovplyvní našu ohodnocovaciu funkciu, ktorá sa pozerá na rok vydania publikácie. Ktorý mesiac a deň vyberieme sa určí pri experimentovaní. Vyberú sa také hodnoty aby veľkosť množiny autorov, bola do 10 000. Taktiež, aby sme ju následne dokázali zobraziť v nástroji Gephi a spracovať v K-Means, v reálnom čase. Takáto podstatná redukcia dát je nutná preto, lebo dátový set je veľmi rozsiahly.

V experimente nám ako najvhodnejší mesiac vyšiel Január a deň Nedeľa.

Experiment v rámci Gephi spočíval v zobrazení grafu a v testovaní rôznych možností rozloženia, filtrov a zobrazení. Ako rozloženie sme použili „Force atlas“. Toto rozloženie je iteratívne a pracuje donekonečna. Preto je ho vhodné zastaviť v momente keď sa body javia ako ustálené. Avšak vždy ho v konečnom dôsledku zastavíme v inom momente. Z toho následne dostaneme vždy iné vzájomné vzdialenosti bodov, ktoré sa používajú pri detegovaní komunít. Taktiež sme experimentovali s rôznymi filtrami a s rôznymi hodnotiacimi faktormi (betweenness centrality a closeness centrality). Ďalej je vhodné nastaviť veľkosť vrcholu podľa počtu prepojení, aby sme zvýraznili významných členov komunít.

V experimente sme zistili, že rozloženie vrcholov je vhodné zastaviť približne po 30 sekundách (pri 10 000 vrcholoch). Ďalej je vhodné vyfiltrovať vrcholy, ktoré majú menej ako 5 prepojení.

Experiment v rámci K-Means spočíval v nastavení parametrov tak, aby sa jeho výsledok približoval „zlatému štandardu“ definovaného podľa vizualizácie v Gephi. Pripomíname, že výsledok K-Means nie je možné vizualizovať (pre N dimenzionálny priestor). Preto sme jeho výstup previedli do textovej formy, kde sme vypísali informácie o klastroch, vrcholoch v rámci klastra a vzdialenosti od centra. Tieto informácie sú priložené v prílohe zadania. Z týchto informácií sme následne vybrali určité vrcholy a pokúsili sa ich porovnať s vizualizáciou v Gephi.

V experimente sme zistili, že najlepšie nastavenie parametrov pre K-Means je:

- Maximálny počet iterácii: 50

- Počet klastrov: 100
- Delta konvergencie: 0.05
- ClusterClassificationThreshold: 0.0

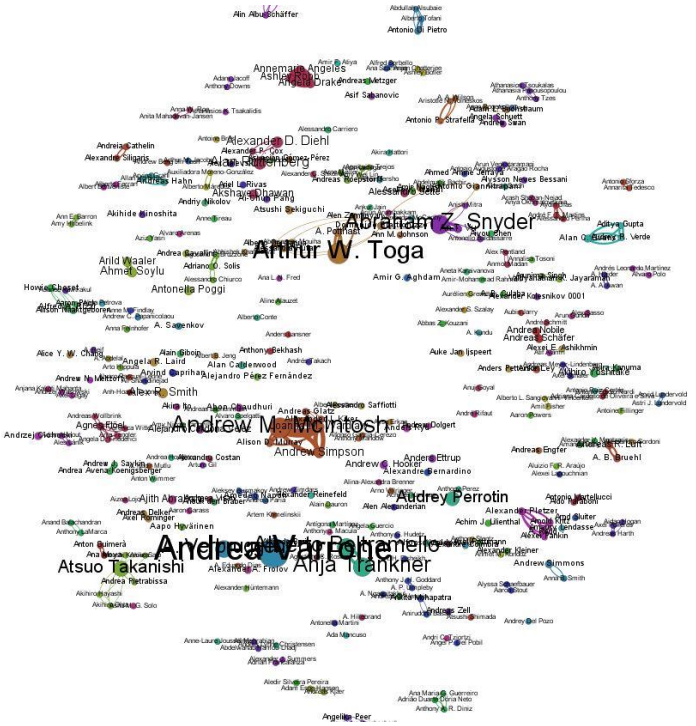

Fig. 2. Objavené komunity pomocou nástroja Gephi.

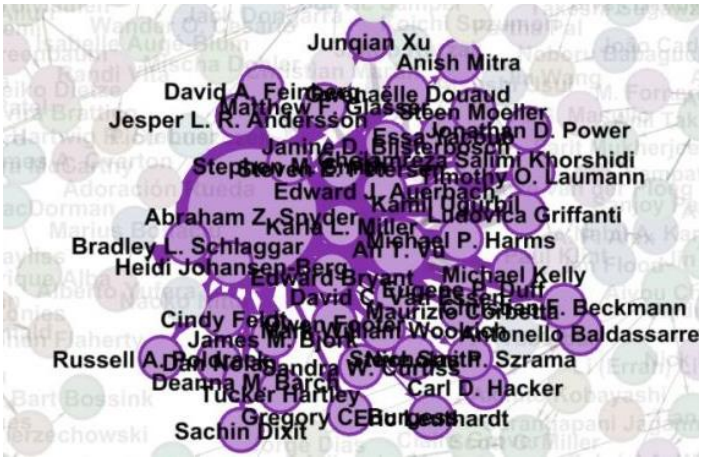

Fig. 3. Detail objavenej komunity.

Taktiež sme sa pokúsili určiť Davies-Bouldin index. Knižnica Mahout tento index už obsahuje, avšak jeho výsledky nám nedávali zmysel a z toho dôvodu sme ho nebrali v úvahu.

8 Vyhodnotenie

V časti DM metódy sme poukázali, že použitie neoverenej metódy nemusí priniesť požadované výsledky. Rovnako použitie implementovanej metódy z vyvíjajúcej knižnice môže priniesť prekvapujúce výsledky. Takže našu navrhovanú metódu Fuzzy K-Means považujeme za pokus, ktorý bol z externých dôvodov označený ako nedokončený a je potrebné ho overiť na spoľahlivej implementácii.

Experiment s nástrojom Gephi nám dal predstavu o množstve komunít, ktoré sa nachádzajú v našich dátach. Následne sme sa snažili tento počet komunít (klastrov) aplikovať aj na algoritmus K-Means. Naším cieľom bolo otestovať, či sa výsledky budú aspoň čiastočne zhodovať s výsledkami s Gephi.

Výsledky nám ukázali, že významní autori nájdení pomocou algoritmu Modularity v nástroji Gephi, sú vo väčšine prípadov taktiež označení ako významní autori aj pomocou nami otestovaného algoritmu K-Means. Algoritmus K-Means ich priraďoval približne do rovnakých klastrov a ohodnotenie významných vrcholov (autorov) bolo vysoké. Rovnako významní autori boli v Gephi zvýraznení prostredníctvom veľkosti vrcholu.

V nami použitej vzorke sme identifikovali veľké množstvo komunít, ktorých početnosť je relatívne malá. Avšak tento výsledok môže byť správny, keď si uvedomíme fakt, akú dátovú vzorku používame. Ide o dáta ktoré sú vyberané v podstate náhodne vzhľadom na celkovú množinu autorov v komunite. Mnohí autori sú vo veľkej miere špecializovaný a publikujú iba v rámci určitej oblasti. Tým sa mohutnosť komunít môže radikálne zmenšiť.

9 Ďalšia práca

Na ďalšiu prácu v detegovaní komunít máme viacero možností. Prvou je skombinovať náš dátový set s nejakým ďalším. Cieľom môže byť jeho rozšírenie, alebo získanie ďalších informácií ktoré nám môžu pomôcť získať kvalitnejšie výsledky napr. H-index. Tento údaj nie je priamo obsiahnutý v dátovom sete DBLP, avšak je ho možne získať z online knižnice CiteSeerX[3].

V mnohých prípadoch je pre používateľa zaujímavé ak sa môže dozvedieť o vizualizovaných autoroch aj viac ako je iba meno. Túto problematiku rieši expertné profilovanie. Kde sa priamo pri každom autorovi udržiavajú informácie o jeho kvalite

[3] http://citeseerx.ist.psu.edu/index

vo všetkých oblastiach v ktorých publikuje, kompletný zoznam publikácií, citované práce, projekty na ktorých sa autor zúčastňoval, organizácie kde pracoval a mnoho ďalšieho.

Jednou z možností je aj vylepšenie ohodnocovacej funkcie. Doposiaľ sa do detegovania komunít započítava počet autorov na publikácií a rok publikácie. Avšak pre dosiahnutie presnejších a kvalitnejších výsledkov je možné vziať v úvahu aj ďalšie informácie. V našej práci sa staticky každému typu publikácie priradí ohodnotenie. Tu je možné vylepšenie, kde by ohodnotenie bolo jedinečné a odzrkadľovalo by konkrétnu kvalitu publikácie. Na to ohodnotenie je možné využiť informácie ako je počet citácií publikácie, konferencia kde bola práca publikovaná atď. Opäť by bolo potrebné ďalšie údaje získať z iných zdrojov napr. počet citácií publikácie z Google Scholar[4], ktorý ich implicitne obsahuje.

Ako ďalšiu prácu je možné definovať aj použitie celej dátovej sady. V našej práci sme prezentovali výsledky, ktoré boli získane iba z určitej vzorky dát. Teda naše výsledky nemajú až takú významnú štatistickú hodnotu.

10 Použitá literatúra

[1] BLONDEL, V.D. et al. Fast unfolding of communities in large networks. In *J. Stat. Mech*. 2008. s. P10008. .

[2] LIU, Q. et al. A new K-means algorithm for community structures detection based on fuzzy clustering. In *Granular Computing (GrC), 2012 IEEE International Conference on*. 2012. s. 1–5. .

[3] OWEN, S. et al. *Mahout in Action* [online]. 1. vyd. [s.l.]: Manning Publications, 2011. ISBN 1935182684.

[4] SANTOS, C.K. DOS et al. Potential Collaboration Discovery Using Document Clustering and Community Structure Detection. In *Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management* [online]. New York, NY, USA: ACM, 2009. s. 39–46. Dostupné na internete: <http://doi.acm.org/10.1145/1651274.1651283>.

[4] http://scholar.google.sk/