Západočeská univerzita v Plzni
Katedra informatiky a výpočetní techniky
Univerzitní 8
306 14 Plzeň

Určování významnosti vrcholů grafu:
PageRank a jeho modifikace

Odborná práce ke státní doktorské zkoušce

Michal Nykl

Technická zpráva č. DCSE/TR-2013-9
Listopad, 2013

Distribuce: veřejná

Technická zpráva č. DCSE/TR-2013-9
Listopad, 2013

Určování významnosti vrcholů grafu:
PageRank a jeho modifikace

Michal Nykl

Abstrakt

V citační analýze existuje mnoho metrik pro měření významu článků, časopisů, autorů atd., jako jsou např. Impact Factor nebo h-index. Tyto metriky ale ve svém výpočtu používají pouze kvantitativní údaje, jako např. počet citací nebo publikací, u kterých nerozlišují kvalitu, tj. při výpočtu je každá citace nebo publikace stejně důležitá. Na základě jejich výsledků lze tedy odlišit populární entity od ostatních, ale již nelze odlišit prestižní entity od populárních entit. Populární entita je často citována, ale prestižní entita je často citována jinými prestižními entitami. K odlišení prestižních entit je v citační analýze stále častěji používán citační graf a algoritmus PageRank, který významnost entity určuje na základě významu entit, které na ní odkazují. Tímto způsobem PageRank svým iteračním výpočtem dokáže odlišit prestižní entity od entit populárních.

Další oblastí, ve které lze PageRank použít, je volba vlastností na základě ontologických vazeb, kde PageRank umožňuje určit významnost vrcholů v grafu tvořeném klíčovými slovy dokumentu. Nejvýznamnější termíny mohou být následně využity k popisu obsahu dokumentu nebo kategorie, do které bude dokument zařazen.

V předkládaném textu jsou krátce shrnuty historie a aktuální stav citační analýzy, popsány algoritmus PageRank a jemu podobné algoritmy a zmíněny používané bibliografické databáze a vytvářené grafy. Dále text obsahuje soupis vlastních dosud dosažených výsledků aplikování PageRanku v oblasti citační analýzy a volby vlastností a popis budoucích výzkumných záměrů.

Kopie zprávy jsou dostupné na
http://www.kiv.zcu.cz/cz/vyzkum/publikace/technicke-zpravy/
nebo na žádost poslanou na následující adresu:

Západočeská univerzita v Plzni
Katedra informatiky a výpočetní techniky
Univerzitní 8
306 14 Plzeň
Česká republika

Obsah

1 Úvod

Cílem tohoto textu je představit téma mé budoucí disertační práce. Úvodní kapitola zmiňuje motivace, které mě vedly k jeho výběru. Druhá kapitola obsahuje aktuální stav poznání v citační analýze, popis algoritmu PageRank a jemu podobných algoritmů a popis bibliografických databází a vytvářených grafů, které se v citační analýze používají. Třetí kapitola shrnuje vlastní dosud dosažené výsledky aplikování PageRanku a obsahuje soupis vlastních publikací. Ve čtvrté kapitole jsou nastíněny cíle disertační práce a navrhovaný postup práce.

1.1 Motivace

Algoritmus PageRank má velký potenciál v určování významných vrcholů grafu díky jeho výpočtu, ve kterém je významnost vrcholu určena v závislosti na významnosti vrcholů, ze kterých na daný vrchol vede hrana. Této vlastnosti lze využít při vyhodnocování grafů vzniklých z dat z různých oblastí výzkumu, činností nebo služeb (jak je krátce zmíněno v části 2.7).

Hlavní oblastí, ve které chceme PageRank zkoumat, je hodnocení autorů vědeckých publikací. Motivací je pro nás schopnost PageRanku odlišit prestižní autory od autorů populárních na základě citačního grafu. Vypočítané pořadí autorů může být použito v situacích, kdy chceme autory porovnávat (např. při výběrovém řízení), oceňovat (např. přidělovat granty nebo udělovat ocenění za výzkum) nebo sledovat vývoj jejich významu pro vědeckou komunitu, či může být použito tam, kde hodnocení aktuálně závisí pouze na lidském úsudku, který je ovlivnitelný. Zde může být podpůrným prostředkem pro rozhodování a doplnit lidský úsudek nebo ho zcela nahradit. Výhodou je, že automatizované hodnocení má jasně stanovená pravidla, která jsou jednotná pro všechny hodnocené subjekty a nedochází tak ke zvýhodňování subjektů, které jsou lidským hodnotitelům blízké či známé. Proto lze říci, že automatizované hodnocení je spravedlivější, než hodnocení lidské. To, že PageRank poskytuje spravedlivé hodnocení autorů (či obecně vrcholů grafu), potvrzuje i z výpočtu vyplývající skutečnost, že význam autora závisí na významnosti autorů, kteří na něj odkazují (tito autoři potvrzují a určují jeho významnost). PageRank tedy využívá i odvozených vlastností (např. význam citujících autorů), na rozdíl od aktuálně užívaných metrik, které využívají pouze kvantitativních vlastností (např. počet publikací nebo citací). Při detailnější analýze lze využít i vypočítaných hodnot PageRanku, které udávají míru významnosti vrcholu. Z rozdílu hodnot PageRanku dvou vrcholů lze určit, jak se vrcholy liší svou významností.

Druhou oblastí, která nás zajímá, je volba vlastností využitím ontologie. Motivací je pro nás možnost použití těchto metod v klasifikaci, shlukování nebo štítkování dokumentů, což může být využito ve vyhledávačích nebo v systémech sloužících pro automatickou správu dokumentů. Výhodou je, že využitím ontologie (v našem případě Linked Data) jsme schopni doplnit v dokumentu nalezená klíčová slova o další termíny, které jsou (obvykle[1]) zobecněním termínů původních. PageRankem následně zjistíme, který termín nebo termíny nejlépe zastupují daný dokument.

Mezi další oblasti, ve kterých bychom chtěli testovat použití PageRanku, patří analýza sociálních a podobných sítí (např. analýza interakce firem, kde bychom chtěli určit firmy, které nejvíce ovlivňují ostatní firmy) a predikce budoucí hrany (např. predikce vítěze sportovního utkání na základě odehraných utkání – zde očekáváme, že PageRank poskytne lepší predikci vítěze, než tabulka s aktuálním umístěním týmů v soutěži nebo tabulka z předešlé sezóny).

[1] Vazby mezi termíny v Linked Data mohou vyjadřovat nadřazený či podřazený pojem nebo synonymum.

2 Aktuální stav poznání

Cílem této kapitoly je seznámit čtenáře s aktuálním stavem poznání v oblasti citační analýzy a výzkumu algoritmu PageRank, který je v citační analýze stále častěji využíván. Historie citační analýzy je zmíněna v části 2.1, v části 2.2 je popsán algoritmus PageRank, jeho výpočet, související problémy a citlivost na změnu parametrů. Část 2.3 obsahuje popis dalších algoritmů, které jsou PageRanku podobné nebo ho upravují. Část 2.4 ukazuje, jakými způsoby mohou být hodnoty např. publikací rozděleny mezi jejich autory. V části 2.5 jsou zmíněny grafy, které lze vytvořit z bibliografických dat a následně vyhodnocovat algoritmem PageRank, nejznámější bibliografické databáze a významná ocenění, která mohou sloužit k porovnání kvality vypočítaných pořadí autorů. Část 2.6 obsahuje popis několika metrik pro porovnání celých grafů a odkaz na několik dostupných programových knihoven, které mohou být při vyhodnocování grafu použity. V poslední části 2.7 jsou zmíněny oblasti výzkumu, ve kterých bylo testováno použití algoritmu PageRank.

2.1 Citační analýza

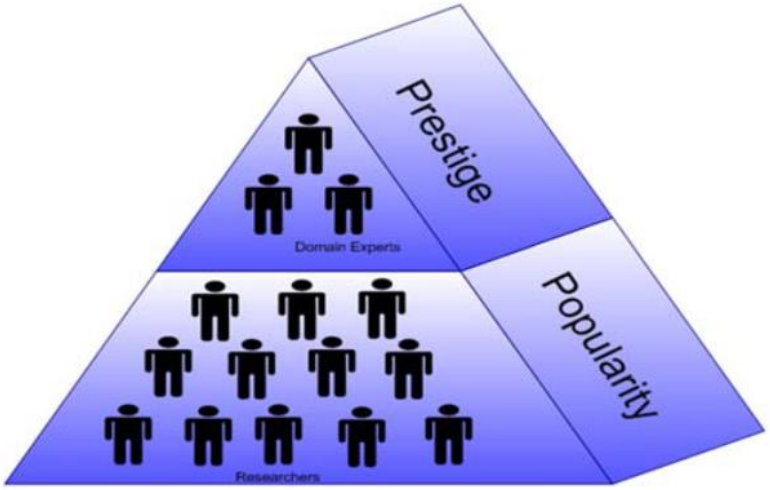
Jedním ze zakladatelů citační analýzy je Eugen Garfield (Garfield, 1955), který jako první navrhl systematické indexování vědecké literatury a citací v ní obsažených za účelem tvorby citačního indexu sloužícímu k hodnocení vědeckých časopisů. Navrženou metodu hodnocení časopisů nazval *Impact Factor*. Cílem návrhu bylo využití Impact Factoru v historickém výzkumu, pro zhodnocení významu určitého vědeckého díla a jeho vlivu na literaturu a myšlení v daném období. Garfield poznamenává, že Impact Factor indikuje vliv časopisů více, než absolutní počet publikací, který již dříve použili Lehman a Dennis (1954). Dále uvádí, že postup je podobný kvantitativní metrice, kterou používal Gross (1944) při hodnocení významu vědeckých časopisů (tato metrika byla však kritizována). V oblasti hodnocení autorů vědeckých publikací lze za nejznámější hodnotící index považovat *h-index* (Hirsch, 2005).

Častým cílem citační analýzy je odlišení populárních a prestižních autorů. V (Y. Ding, 2011) autorka odkazuje na skutečnost, že pojem *populární* vychází z latinského výrazu *popularis*[2], kterému lze rozumět jako „*být milovaný lidmi*“. Kdežto pojem *prestižní*, z latinského *praestigious*[3], lze chápat jako „*mající oslnivý vliv*“. Nástin odlišení populárních a prestižních autorů můžeme vidět na obr. 1. V bibliometrii lze říci, že populární autor je hodně citovaný a popularitu tedy lze měřit *počtem citací*. Naopak prestižní autor je citovaný významnými autory a prestiž tedy lze měřit *počtem citací od významných autorů* (což znamená, že je potřeba určit, kdo je významný a kdo není). Autor může být populární, ale nemusí být prestižní a naopak. Autorka uvádí pěkný příklad, když říká, že autor, který ve své práci shrnuje aktuální stav poznání v určité oblasti, může být hodně citován začínajícími autory v dané oblasti, ale již méně těmi, kteří jsou v dané oblasti experty – autor je populární. Naopak autor referátu, který představuje inovativní metodu, může být citován experty, ale již méně laiky – autor je prestižní.

S ohledem na výše uvedené odlišení pojmů *populární* a *prestižní* se v citační analýze pozvolna přechází od metrik používajících pouze kvantitativní vlastnosti (např. počet citací atd.) k metrikám používajícím i odvozené vlastnosti. Tyto metriky využívají významnosti citujících entit a tím dokáží určit, zda citace pochází z prestižního zdroje, a např. ji zvýhodnit. Za tímto účelem se v citační analýze stále častěji používá algoritmus PageRank (Y. Ding, 2011) či jeho úpravy. Cílem citační analýzy je tedy

[2] Výklad slova „populární“ - http://www.etymonline.com/index.php?term=popular
[3] Výklad slova „prestižní“ - http://www.etymonline.com/index.php?term=prestigious

nalezení významných či prestižních autorů (popř. článků, časopisů, institucí, témat atd.) využitím algoritmů nebo metod, které pracují s bibliografickými záznamy a citačním grafem. Tento problém lze algoritmicky zapsat takto: na vstupu máme bibliografické záznamy o publikacích z určité vědecké oblasti (např. počítačové vědy) a na výstupu chceme získat hodnoty významnosti autorů těchto publikací, dle kterých můžeme autory seřadit.



Obr. 1: Prestižní (citovaný hodně citovanými články) a populární (citovaný normálními články). Obrázek přejat z (Y. Ding, 2011).

Využitím citační analýzy lze určovat významnost časopisů a následně dle ní vybírat časopisy do vědeckých knihoven či bibliografických databází nebo vybírat časopisy, ve kterých bychom chtěli publikovat své vědecké výsledky. Se stejným záměrem můžeme využít vypočítané významnosti konferencí. Publikace mohou být také vyhodnocovány s cílem určení jejich významnosti, či pro zjištění jejich vědeckého přínosu. Významnost výzkumných institucí lze využít při rozdělování finančních prostředků, kde jí můžeme zahrnout do státního systému hodnocení výzkumných institucí, nebo při vytváření jejich pořadí. Také témata mohou být vyhodnocena citační analýzou. Zde se obvykle ptáme, které téma je nejvíce rozvíjené či přínosné. Využitím citační analýzy lze také vytvářet pořadí států a porovnávat tak jejich přínos k celosvětovému vědeckému rozvoji.

Jako relevantní zdroje lze uvést např.: hodnocení časopisů (Bollen et al., 2006; Garfield, 1972), konferencí (Sidiropoulos & Manolopoulos, 2005b), publikací (Sidiropoulos & Manolopoulos, 2005a; Siler, 2012), institucí (Ho, 2013; Mryglod et al., 2013), témat (H. Wang et al., 2012) a států (Leydesdorff, 2013; Ma et al., 2008). Tento výčet relevantních zdrojů samozřejmě není kompletní.

Více základních informací týkajících se citační analýzy lze nalézt např. v (Bellis, 2009; Moed, 2005).

2.2 Algoritmus PageRank

Algoritmus PageRank (Brin & Page, 1998; Page et al., 1999) byl vyvinut pro určení významností webových stránek, které se následně využívají při řazení relevantních stránek ve výsledcích vyhledávačů, např. Google.com. Jak autoři uvádějí (Page et al., 1999), jeho koncept vychází z citační analýzy. Při určování významnosti webové stránky se využívá hypertextových odkazů, které na stránku odkazují, a významností stránek, ze kterých odkazy vedou (algoritmus je iterativní). Z matematického hlediska je vyhodnocován graf, jehož vrcholy jsou webové stránky a hrany vyjadřují,

že z jedné webové stránky vede hypertextový odkaz na stránku jinou. Pokud tento graf navíc splňuje definici (Ryjáček, 2001): *„Síť je orientovaný graf s kladným reálným ohodnocením hran a s reálným (připouštíme i záporné hodnoty) ohodnocením uzlů.“*, tak lze hovořit o síti. Interní hypertextové odkazy (lze použít označení z bibliografie „samocitace“), tj. ty odkazy, které odkazují na stránku, na které se nacházejí, se při vyhodnocování Webu nepoužívají. Měli bychom také zmínit, že PageRank je aplikací Markovova řetězce (Langville & Meyer, 2006).

2.2.1 Matematický zápis

Algoritmus PageRank lze popsat buďto maticovým zápisem, který je užitečný pro matematické zkoumání algoritmu (např. jeho konvergence, urychlení výpočtu atd.), nebo zápisem výpočtu pro jeden prvek, což je užitečné pro jeho „snazší“ pochopení a implementaci. Dále si ukážeme oba dva zápisy, přičemž využijeme maticové zápisy a důkazy uvedené v (Langville & Meyer, 2006).

Prvním vzorcem PageRanku, tak jak ho navrhli Page a Brin v (Page et al., 1999), je vzorec (1), kde $PR_x(A)$ je skóre PageRanku vrcholu A v iteraci x, U je množina všech vrcholů odkazujících na vrchol A a N_u je počet výstupních hran vrcholu u. Tomuto zápisu odpovídá maticový zápis uvedený ve vzorci (2), kde $\boldsymbol{\pi}^{(x)}$ je vektor PageRanků všech vrcholů grafu v iteraci x a $\boldsymbol{H}$ je řádkově normalizovaná matice sousednosti.

$$PR_{x+1}(A) = \sum_{u \in U} \frac{PR_x(u)}{N_u} \tag{1}$$

$$\boldsymbol{\pi}^{(x+1)T} = \boldsymbol{\pi}^{(x)T}\boldsymbol{H} \tag{2}$$

Hodnota PageRanku udává pravděpodobnost, s jakou se webový surfař, pohybující se po Webu využíváním hypertextových odkazů, dostane na danou webovou stránku. Součet PageRanků všech vrcholů grafu je roven 1, tj. 100%.

Prvním problémem, se kterým se vzorce (1) a (2) potýkají, je problém slepých vrcholů (*dangling nodes)*, tj. vrcholů, které nemají žádné výstupní hrany. V těchto vrcholech dochází ke ztrátě hodnoty PageRanku a součet PageRanků všech vrcholů grafu tak přestává být roven 1. Lze říci, že surfaři, kteří jsou aktuálně ve slepých vrcholech, budou v následující iteraci kdesi „mimo graf“. Tento problém lze řešit třemi způsoby:

1) vytvoření stoku - vytvoříme v grafu nový vrchol (stok) se samocitační hranou (smyčkou) a všem slepým vrcholům přidáme výstupní hranu směřující na tento vrchol;
2) normalizace – po každé iteraci normalizujeme hodnoty PageRanku všech vrcholů grafu tak, aby jejich součet byl roven jedné;
3) rovnoměrné rozdělení – každému slepému vrcholu přidáme výstupní hrany na všechny vrcholy grafu (i na sebe sama).

První způsob (vytvoření stoku) není příliš vhodný, jelikož může vést k situaci, kdy nově vytvořený vrchol získá veškerou hodnotu PageRanku. Druhý způsob (normalizace) není „spravedlivý“, protože každému vrcholu je normalizací přidána jiná hodnota. Jelikož první ani druhý způsob nejsou ideální, používá se (obvykle) způsob třetí. Výstupní hrany ze slepých vrcholů však nemusíme doplňovat přímo do grafu, ale stačí s nimi pouze počítat, jak je ukázáno ve vzorci (3), kde D je množina všech slepých

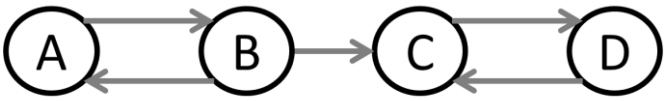
vrcholů grafu, V je množina všech vrcholů grafu a $|V|$ její velikost. Hodnota předávaná od slepých vrcholů se v průběhu iterace nemění a lze ji vypočítat na začátku každé iterace, čímž lze výpočet urychlit.

$$PR_{x+1}(A) = \sum_{u \in U} \frac{PR_x(u)}{N_u} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \tag{3}$$

Pokud bychom tuto úpravu chtěli zapsat maticovým zápisem, je nejsnazší vytvořit matici $\boldsymbol{S}$, dle vzorce (4), a nahradit s ní matici $\boldsymbol{H}$ ve vzorci (2). Ve vzorci (4) je $\boldsymbol{a}$ vektor slepých vrcholů, kde a_i je rovno jedné, pokud je vrchol i slepým vrcholem, jinak je a_i rovno nule, n je počet vrcholů grafu a $\boldsymbol{e}$ je jednotkový vektor.

$$\boldsymbol{S} = \boldsymbol{H} + \boldsymbol{a}\left(\frac{1}{n}\boldsymbol{e}^T\right) \tag{4}$$

Vzorce (1) až (4) ovšem neřeší problém zvaný *Rank sink* (volně přeloženo: stok hodnot), který vzniká, pokud vrcholy ve skupině odkazují sami na sebe, ale neodkazují vně skupiny, přičemž skupina je odkazována z vnější. Rank sink ilustruje obr. 2 s příkladem, ve kterém vrcholy A a B při výpočtu PageRanku postupně předají celý svůj PageRank vrcholům C a D a PageRank vrcholů A a B bude (po dostatečném množství iterací) roven nule. Dalším problémem je, že pokud vrcholy C a D nebudou mít každý přesně polovinu celkového PageRanku, tak si v každé iteraci vymění své PageRanky a nikdy nenastane ustálený stav, tj. algoritmus nebude konvergovat.



Obr. 2: příklad ilustrující klesání hodnocení při výpočtu PageRanku.

Pro řešení tohoto problému autoři navrhli model náhodného surfaře, který se Webem pohybuje klikáním na hypertextové odkazy nebo využívá tzv. teleportu, když přejde na náhodnou stránku tak, že do webového prohlížeče přímo zadá její URL. Autoři, sledováním chování reálných uživatelů internetu, zjistili, že teleportu uživatelé využívají cca jednou za 7 kroků. Proto i ve svém algoritmu stanovili užití teleportu s pravděpodobností 15% (Brin & Page, 1998). Do algoritmu PageRank byl model náhodného surfaře vložen konstantou *d* zvanou faktor tlumení. Surfař tak s pravděpodobností *d* následuje hypertextové odkazy nebo s pravděpodobností *1-d* využije teleportu. Faktor tlumení je tedy obvykle nastaven na hodnotu 0,85. Hodnota ale může být měněna, pokud chceme klást větší důraz na využití hypertextových odkazů, či na personalizaci (neuniformní úprava náhodného teleportu). Čím blíže jedné faktor tlumení je, tím více iterací, k dosažení zvolené přesnosti výsledků, je potřeba (Langville & Meyer, 2006; Nykl, 2011).

Vzorec (5) ukazuje, jak byl vzorec (3) doplněn o faktor tlumení a hodnota, kterou každý vrchol získá díky „teleportu“, normalizována.

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot \left(\sum_{u \in U} \frac{PR_x(u)}{N_u} + \frac{1}{|V|} \sum_{s \in D} PR_x(s)\right) \tag{5}$$

Pokud bychom stejnou úpravu chtěli provést v maticovém zápisu, vytvoříme matici ***G*** dle vzorce (6) a nahradíme s ní matici ***H*** ve vzorci (2).

$$\boldsymbol{G} = \frac{(1-d)}{n}\boldsymbol{e}\boldsymbol{e}^T + d\boldsymbol{S} \tag{6}$$

Poslední nepřesností vzorce (5), která se nám ovšem v prostředí Webu obvykle neprojeví, je, že každý hypertextový odkaz má ve výpočtu stejnou váhu, tj. jsou-li na stránce např. 4 odkazy, uvažujeme, že každý z nich bude použit s pravděpodobností ¼. Pokud bychom chtěli některý z odkazů zvýhodnit, museli bychom vzorec (5) doplnit o váhy hran, jak ukazuje vzorec (7), kde w_{utoA} je váha hrany vedoucí z vrcholu *u* do vrcholu *A* a $w_{u\text{out}}$ je součet vah všech výstupních hran vrcholu *u*.

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot (\sum_{u \in U} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|}\sum_{s \in D} PR_x(s)) \tag{7}$$

V maticovém zápisu žádnou úpravu, přidávající do výpočtu váhy hran, provádět nemusíme, jelikož váhy hran můžeme zapsat do matice sousednosti ***H*** a tu následně řádkově normalizovat.

S odkazem na matematické důkazy v (Langville & Meyer, 2006) můžeme říci, že výpočet PageRanku konverguje k jedinečnému výsledku bez ohledu na počáteční hodnoty PageRanků. Přesto se ale obvykle počáteční hodnoty vrcholů nastavují na hodnotu *1/|V|* nebo na hodnoty uvedené v personalizačním vektoru, či nejlépe na hodnoty blízké konečnému výsledku.

2.2.2 Personalizace

V některých případech požadujeme, aby některé vrcholy byly algoritmem PageRank v průběhu výpočtu zvýhodněny. Toho můžeme docílit neuniformním rozdělením faktoru tlumení (přesněji rozdělením *1-d*), čemuž říkáme *personalizace*. Pojem zavedli samotní autoři (Page et al., 1999), když uvažovali jak do výpočtu zahrnout různé potřeby či vlastnosti uživatelů. Vzorec PageRanku doplněný o personalizaci znázorňuje vzorec (8), kde *P* je vektor personalizací všech vrcholů grafu a P_A je hodnota personalizace vrcholu *A*.

$$PR_{x+1}(A) = \frac{(1-d) \cdot P_A}{\sum_{p \in P} p} + d \cdot (\sum_{u \in U} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|}\sum_{s \in D} PR_x(s)) \tag{8}$$

Ve vzorci (8) předpokládáme, že nenastane situace, kdy $\sum p = 0$. Pokud by tato situace mohla nastat, můžeme využít vzorec (9). Ovšem možností řešení je více a vždy záleží na konkrétním použití algoritmu.

$$PR_{x+1}(A) = \frac{(1-d) \cdot (1+P_A)}{|V| + \sum_{p \in P} p} + d \cdot (\sum_{u \in U} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|}\sum_{s \in D} PR_x(s)) \tag{9}$$

Vzorci (8) odpovídá maticový zápis vzorce (10) pro výpočet matice $\boldsymbol{G}$, kde $\boldsymbol{v}$ je personalizační vektor. Pro výpočet hodnot PageRanků se matice $\boldsymbol{G}$ opět použije ve vzorci (2) místo matice $\boldsymbol{H}$.

$$\boldsymbol{G} = (1 - d)\boldsymbol{e}\boldsymbol{v}^T + d\boldsymbol{S} \tag{10}$$

Pěkný příklad použití personalizace v citační analýze můžeme nalézt v (Yan & Ding, 2011), kde autoři pomocí PageRanku vyhodnocují graf spoluautorství, a navíc používají personalizaci tak, že P_A představuje počet citací, které obdržel autor *A*. Použitím faktoru tlumení o velikosti 0,55 tak autoři kombinují hodnoty získané z grafu spoluautorství a hodnoty získané z citačního grafu autorů.

2.2.3 Citlivost na změnu parametrů

Pokud bychom chtěli zkoumat, jak se změní výsledek PageRanku při změně některého parametru, měli bychom začít faktorem tlumení. Informace jsou přejaty z (Langville & Meyer, 2006).

Pokud je faktor tlumení malý, tak výsledek není příliš citlivý na jeho malou změnu. Čím větší ale faktor tlumení je, tím je výsledek citlivější na jeho malou změnu, a je-li faktor tlumení blízký jedné, je výsledek na jeho malou změnu velmi citlivý.

Citlivost na změnu matice $\boldsymbol{H}$ (či matic $\boldsymbol{S}$ a $\boldsymbol{G}$) opět závisí na velikosti faktoru tlumení. Je-li faktor tlumení blízký jedné, je výsledek velmi citlivý na změny ve struktuře grafu. Čím blíže je faktor tlumení nule, tím méně je výsledek citlivý na změnu struktury grafu.

Citlivost na změnu personalizačního vektoru závisí také na velikosti faktoru tlumení, ale s opačným efektem. Čím blíže nule faktor tlumení je, tím více je výsledek citlivý na změnu personalizačního vektoru. Je-li faktor tlumení blízký jedné, tak výsledek není příliš citlivý na změnu personalizačního vektoru.

Velikostí faktoru tlumení tedy můžeme regulovat, zda se ve výsledku více zohlední struktura grafu (vysoký faktor tlumení) nebo personalizace (nízký faktor tlumení).

2.3 PageRanku podobné metody pro měření významnosti

V této části jsou zmíněny algoritmy, které vychází z algoritmu PageRank (tj. upravují nebo rozšiřují ho), algoritmy jim podobné a metody hodnocení bibliografických entit, které je využívají.

2.3.1 Vážený PageRank a AuthorRank

Od vzniku PageRanku bylo navrženo několik jeho variant, které autoři označili jako vážený PageRank (*Weighted PageRank* - WPR). Mezi jednodušší varianty patří WPR prezentovaný v (Y. Ding, 2011), kde autoři používají pouze personalizaci a nepoužívají váhy hran. Zato říkají, že personalizace vrcholu, který zastupuje autora, může obsahovat počet citací, počet publikací, počet publikací, kde byl autor uveden jako první autor nebo h-index. Autoři následně používají počet citací a počet publikací a testují faktor tlumení 0,15, 0,5 a 0,85. V předchozím článku (Y. Ding & Yan, 2009) autoři vyhodnocují co-citační graf autorů a používají stejné dvě varianty personalizačního vektoru. Měli bychom zmínit, že klasickou variantou WPR je vzorec (8), který uvažuje váhy hran i personalizaci. Variantu PageRanku, která uvažuje váhy hran a neuvažuje personalizaci, použitou na graf spoluautorství, nazvali autoři *AuthorRank* (Liu et al., 2005).

Další variantu, kterou autoři označují jako WPR, lze nalézt v (Xing & Ghorbani, 2004), kde autoři říkají, že významný vrchol je hodně provázaný s ostatními a ostatní vrcholy chtějí na tento vrchol odkazovat a chtějí jím být odkazovány. Proto ve svém vzorci uvažují vstupní i výstupní hrany vrcholu.

2.3.2 Bibliografický PageRank a Time-aware PageRank

Označení bibliografický PageRank použili autoři (Fiala et al., 2008) pro pojmenování jejich úprav PageRanku pro vyhodnocení citačního grafu autorů. Autoři nejprve definují bipartitní graf autorství (autoři a publikace) a citační graf publikací a s jejich využitím konstruují citační graf autorů. V tomto grafu vede mezi dvěma autory A a B orientovaná hrana, pokud autor A ve své publikaci citoval publikaci autora B a navíc žádná z obou publikací není napsána oběma autory. Autoři následně používají upravený vzorec PageRanku (11), kde $R_x(u)$ je PageRank vrcholu u v iteraci x, d je faktor tlumení, A je množina všech vrcholů grafu (autorů), E je množina všech hran v citačním grafu autorů, (v,u) je hrana vedoucí z vrcholu v do vrcholu u a $\sigma_{v,u}$ je konstantní váha přiřazená hraně (v,u) dle vzorce (12).

$$R_{x+1}(u) = \frac{(1-d)}{|A|} + d \cdot \sum_{(v,u)\in E} \frac{PR_x(v) \cdot \sigma_{v,u}}{\sum_{(v,k)\in E} \sigma_{v,k}} \tag{11}$$

$$\sigma_{v,k} = \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \cdot \sum_{(v,j)\in E} w_{v,j}} \tag{12}$$

Ve vzorci (12) je $w_{v,k}$ počet hran (citací) vedoucích z vrcholu v do vrcholu k, $c_{v,k}$ je počet společných publikací autorů v a k a $b_{v,k}$ je jednou z následujících sedmi variant: **a)** nula; **b)** počet publikací obou autorů; **c)** počet všech spoluautorů autora v plus počet všech spoluautorů autora k (každý spoluautor se počítá tolikrát, kolikrát byl danému autoru spoluautorem); **d)** počet unikátních spoluautorů autora v plus počet unikátních spoluautorů autora k (každý spoluautor je pro každého autora počítán pouze jednou); **e)** počet publikací autora v plus počet publikací autora k, přičemž se počítají pouze ty publikace, které obsahují alespoň jednoho spoluautora; **f)** počet všech spoluautorů ve společných publikacích autorů v a k (každý spoluautor je počítán tolikrát, kolikrát byl spoluautorem); **g)** počet unikátních spoluautorů ve společných publikacích autorů v a k (každý spoluautor je počítán pouze jednou).

Autoři se snaží penalizovat citovaného autora frekvencí spolupráce s citujícím autorem. Říkají, že citace od spoluautora je méně významná, než citace od cizího autora. Ve svém experimentu následně používají data DBLP a výsledky porovnávají s *ACM SIGMOD E.F. Codd Innovations Award*. Jako nejlepší varianty jim vychází varianty d) a e) těsně následované variantami b) a f), přičemž varianty b), d) a e) spolu vysoce korelují (použit byl Spearmanův koeficient korelace).

V následující práci (Fiala, 2012b) autor přidává k výše zmíněným variantám čas publikování a vzniku citace a vytváří tak novou variantu PageRanku, kterou nazývá *Time-aware PageRank*. $c_{v,k}$ i $b_{v,k}$ tak získají horní index t značící rok, do kterého se daná veličina počítá, tj. $c^t_{v,k}$ představuje počet společných publikací autorů v a k před rokem t. Vzorec PageRanku je upraven na vzorce (13) a (14).

$$R_{x+1}(u) = \frac{(1-d)}{|A|} + d \cdot \sum_{(v,u)\in E} \frac{PR_x(v) \cdot \sum_{i=1}^{w_{v,u}} \sigma_{v,u}^i}{\sum_{(v,k)\in E} \sum_{i=1}^{w_{v,k}} \sigma_{v,k}^i} \tag{13}$$

$$\sigma_{v,k}^i = \frac{1}{\frac{c_{v,k}^i + 1}{b_{v.k}^i + 1} \cdot \sum_{(v,j)\in E} 1} \tag{14}$$

Experiment s Time-aware PageRankem autor provádí na datech WoS, kde testuje varianty čas uvažující i neuvažují, a ukazuje, že, v porovnání s *ACM A.M. Turing Award* a *ACM SIGMOD E.F. Codd Innovations Award*, nejlepších výsledků dosahují varianty d) a c) uvažující čas.

2.3.3 HITS

Výpočet HITS (*Hypertext Induced Topic Search*) (Kleinberg, 1999) je podobný PageRanku, ovšem se dvěma zásadními rozdíly. PageRank není závislý na dotazu, kdežto HITS je na dotazu závislý, tj. PageRank se počítá pro celý webový graf a jeho hodnoty jsou do vyhledávání zařazeny až následně, kdežto HITS se počítá nad grafem vytvořeným z výsledků vyhledávání, které se s jeho pomocí seřadí. Druhým rozdílem je, že PageRank uvažuje, že všechny vrcholy grafu jsou stejného typu a každému vrcholu přiřazuje jednu hodnotu. HITS uvažuje u každého vrcholu dvě vlastnosti a každému vrcholu přiřazuje dvě hodnoty. HITS pomyslně dělí vrcholy na autority (*authorities*) a rozcestníky (*hubs*) a říká, že: „*dobré rozcestníky jsou ty vrcholy, které odkazují na dobré autority, a dobré autority jsou ty vrcholy, které jsou odkazované dobrými rozcestníky*“. V duchu této kruhové definice je HITS i počítán.

Vzorec (15) znázorňuje výpočet skóre autorit a vzorec (16) výpočet skóre rozcestníků, kde $\boldsymbol{x}^{(k)}$ je vektor skóre autorit v iteraci *k*, $\boldsymbol{y}^{(k)}$ je vektor skóre rozcestníků v iteraci *k* a $\boldsymbol{L}$ je matice sousednosti.

$$\boldsymbol{x}^{(k)} = \boldsymbol{L}^T \boldsymbol{y}^{(k-1)} \tag{15}$$

$$\boldsymbol{y}^{(k)} = \boldsymbol{L}\boldsymbol{x}^{(k)} \tag{16}$$

Vzorce (15) a (16) lze zbavit kruhové závislosti, jak je ukázáno ve vzorcích (17) a (18).

$$\boldsymbol{x}^{(k)} = \boldsymbol{L}^T \boldsymbol{L}\boldsymbol{x}^{(k-1)} \tag{17}$$

$$\boldsymbol{y}^{(k)} = \boldsymbol{L}\boldsymbol{L}^T \boldsymbol{y}^{(k-1)} \tag{18}$$

Více o konvergenci, citlivosti, urychlení výpočtu a silných a slabých stránkách algoritmu HITS lze nalézt v (Langville & Meyer, 2006). V článku (C. Ding et al., 2002) autoři kombinují algoritmy PageRank a HITS a ukazují provázanost algoritmu HITS a citační analýzy. Implementaci algoritmu HITS používá vyhledávač Teoma[4].

[4] Vyhledávač Teoma používá HITS pro řazení výsledků vyhledávání - http://www.teoma.com

2.3.4 FutureRank

Algoritmus *FutureRank* (Sayyadi & Getoor, 2009) kombinuje PageRank a HITS. Iterativní výpočet obsahuje jeden krok PageRanku pro výpočet skóre publikací a jeden krok HITS pro výpočet skóre autorů. Skóre publikací je počítáno využitím PageRanku podobného vzorce (19), který obsahuje tři faktory tlumení – faktor α tlumí vliv přenosu skóre mezi publikacemi (užívá se citační graf publikací, M^C je matice citovanosti publikací a R^P je vektor skóre publikací), faktor β tlumí vliv skóre předaného od autorů (užívá se bipartitní graf autorství, M^A je matice autorství a R^A je vektor skóre autorů) a faktor γ tlumí vliv stáří publikace (R^{Time} je vektor, jehož hodnoty se snižují dle stáří publikace, a n je počet publikací). Mezi publikacemi a autory se skóre přenáší podobně jako v algoritmu HITS. Skóre autorů R^A je počítáno dle vzorce (20). Vektor stáří publikací R^{Time} je počítán využitím vzorce (21), kde $T_{current}$ je současný čas, či čas dotazu (pokud je vzorec použit ve vyhledávání), T_i je čas publikování publikace i a rozdíl $T_{current} - T_i$ udává počet let od publikování. Hodnota ρ byla experimentálně stanovena na 0,62.

$$R^P = \alpha * M^C * R^P + \beta * M^{A^T} * R^A + \gamma * R^{Time} + (1 - \alpha - \beta - \gamma)/n \tag{19}$$

$$R^A = M^A * R^P \tag{20}$$

$$R_i^{Time} = e^{-\rho*(T_{current} - T_i)} \tag{21}$$

Pro výpočet skóre publikací autoři použili několik variant FutureRanku. Variantu s nenulovými konstantami α, β, γ, variantu FutureRank(CT), která nevyužívá skóre autorů (β=0), variantu FutureRank(CA), která nevyužívá stáří publikací (γ=0), a variantu značenou jako PageRank (α=0,9; β=0; γ=0). Autoři říkají, že varianta FutureRank(CT) je podobná algoritmu *CiteRank* (Walker et al., 2006) a varianta FutureRank(CA) je podobná algoritmu *CoRank* (Zhou et al., 2007), který také současně hodnotí publikace a autory.

2.3.5 SALSA

Algoritmus SALSA (*the Stochastic Approach for Link-Structure Analysis*) (Lempel & Moran, 2001, 2000) je kombinací algoritmů HITS a PageRank pro účely vyhledávání. SALSA, obdobně jako HITS, počítá skóre autorit a rozcestníků a navíc z PageRanku přejímá koncept Markovova řetězce. Místo matice sousednosti SALSA vytváří bipartitní neorientovaný graf složený z množiny autorit (vrcholy s alespoň jednou vstupní hranou), množiny rozcestníků (vrcholy s alespoň jednou výstupní hranou) a množiny hran, přičemž vrchol se může nacházet v obou množinách vrcholů a hrany nikdy nevedou mezi prvky stejné množiny.

SALSA vytváří dva Markovovy řetězce (Langville & Meyer, 2006) – Markovův řetězec rozcestníků s maticí $\boldsymbol{H}$ a Markovův řetězec autorit s maticí $\boldsymbol{A}$. $\boldsymbol{L}$ je matice sousednosti. $\boldsymbol{L_r}$ je řádkově normalizovaná matice $\boldsymbol{L}$ a $\boldsymbol{L_c}$ je sloupcově normalizovaná matice $\boldsymbol{L}$. Matice $\boldsymbol{H}$ a $\boldsymbol{A}$ obsahují nenulové sloupce a řádky z matic $\boldsymbol{H'}$ a $\boldsymbol{A'}$ vytvořených dle vzorců (22).

$$\begin{aligned} \boldsymbol{H'} &= \boldsymbol{L_r L_c^T} \\ \boldsymbol{A'} &= \boldsymbol{L_c^T L_r} \end{aligned} \tag{22}$$

Skóre rozcestníků a autorit je počítáno zvlášť pro každou souvislou komponentu $\boldsymbol{C}$ matic $\boldsymbol{H}$ a $\boldsymbol{A}$ dle vzorce (23), kde $\boldsymbol{\pi}^{(k)}(\boldsymbol{C})$ je skóre rozcestníků nebo autorit komponenty $\boldsymbol{C}$ v iteraci k. Globální skóre rozcestníků nebo autorit je dáno poměrným sloučením příslušných komponent, tj. pokud matice H měla 5 prvků a byla tvořena dvěma komponentami o 2 a 3 prvcích, tak výsledný vektor se skóre rozcestníků bude obsahovat prvky první komponenty vynásobené 2/5 a prvky druhé komponenty vynásobené 3/5.

$$\boldsymbol{\pi}^{(k+1)T}(\boldsymbol{C}) = \boldsymbol{\pi}^{(k)T}\boldsymbol{C} \tag{23}$$

2.3.6 SCEAS

Algoritmus SCEAS (Sidiropoulos & Manolopoulos, 2005b), součást stejnojmenného systému pro hodnocení vědeckých kolekcí (*Scientific Collection Evaluator by using Advanced Scoring*), svou koncepcí vychází z algoritmu PageRank, který rozšiřuje. První zmínky o algoritmu SCEAS lze nalézt v (Sidiropoulos & Manolopoulos, 2005a), kde autoři zmiňují dvě varianty, které se liší pouze různým nastavením svých proměnných. Poprvé byl SCEAS použit pro hodnocení publikací, jejich kolekcí (konferencí, časopisů, knih) a autorů v citačních grafech vytvořených z dat DBLP.

SCEAS znázorňuje vzorec (24), kde $S_j^{(k)}$ je SCEAS skóre vrcholu j v iteraci k, d je faktor tlumení, U je množina vrcholů, z nichž vede hrana na vrchol j, N_u je počet výstupních hran vrcholu u (váhy hran se zde neuvažují), b je faktor prosazení přímého citovaní a a faktor rychlosti, se kterou prosazení nepřímého citování konverguje k nule. Autoři používají a=e (Eulerovo číslo). Lze říci, že změna skóre vrcholu i ovlivní skóre vrcholu j, který je x-tým vrcholem v řadě (mezi vrcholy i a j je x-1 vrcholů), s faktorem a^{-x}.

$$S_j^{(k+1)} = (1-d) + d \cdot \sum_{u \in U} \frac{S_u^{(k)} + b}{N_u} a^{-1} \tag{24}$$

Pomineme-li, že část *(1-d)* není normalizována počtem vrcholů grafu, tak pokud b=0 a a=1, tak vzorec (24) odpovídá vzorci PageRanku (5), který neošetřuje slepé vrcholy (tj. bez posledního sčítance), či vzorci (7). Ve variantě SCEAS1 autoři používají d=1 a b=1 a ve variantě SCEAS2 používají d=0,85 a b=0.

Jako výhody SCEAS autoři uvádějí (Sidiropoulos & Manolopoulos, 2005a), že výpočet skóre je více ovlivněn počtem vstupních hran, výpočet algoritmu a jeho konvergence jsou velmi rychlé (v porovnání s PageRankem a HITS) a výpočet je méně citlivý na přidání nového vrcholu do grafu.

2.3.7 Citation Count (CC), Balanced CC, B-HITS, B-SALSA a varianty SCEAS

Autoři algoritmu SCEAS ve své další práci (Sidiropoulos & Manolopoulos, 2006) zkoumají různé varianty hodnocení bibliografických entit, které vychází z algoritmů PageRank, HITS, SALSA a SCEAS. Nejprve definují *Citation Count* (CC – počet citací) jako počet vstupních hran vrcholu a následně *Balanced Citation Count* (BCC – vyvážený počet citací) jako součet částí, které vrchol získá od vrcholů, které ho citují, jak ukazuje vzorec (25), kde BCC_x je Balanced Citation Count vrcholu x, U je množina vrcholů citujících vrchol x a N_u je počet výstupních hran vrcholu u. Váhy hran se neuvažují. CC i BCC autoři kritizují, protože se při jejich výpočtu nepoužívá významnost vrcholů.

$$BCC_x = \sum_{u \in U} \frac{1}{N_u} \tag{25}$$

Autoři následně ukazují PageRank, HITS a SALSA a zavádí míru *Prestiž*, kterou definují jako součet Prestiží citujících vrchol, jak ukazuje vzorec (26), kde P_x je Prestiž vrcholu x a U je množina vrcholů citujících vrchol x. Zde ovšem namítají, že Prestiž vrcholů neúčastnících se žádného cyklu konverguje k nule. Dále, pokud existuje v grafu cesta, tak Prestiž vrcholu x, který cituje vrchol y, nebude nikdy větší než Prestiž vrcholu y. Mimo tyto problémy je s Prestiží spojeno několik dalších problémů, které byly řešeny v návrhu algoritmu PageRank, např. rank sink (viz část 2.2).

$$P_x = \sum_{u \in U} P_u \tag{26}$$

Všechny zmíněné metody autoři kritizují. PageRank kritizují, protože vrcholy účastnící se cyklů získají největší skóre. SALSA a HITS kritizují, protože uvažují rozcestníky a autority, což příliš neodpovídá hodnocení publikací. Proto zavádí jejich úpravy.

Balanced HITS či B-HITS při výpočtu skóre autorit vedle hran vedoucích z rozcestníků uvažuje i hrany vedoucí z jiných autorit. To ukazuje vzorce (27), kde BHA_x je B-HITS autoritové skóre vrcholu x, BHH_x je B-HITS rozcestníkové skóre vrcholu x, U je množina vrcholů, z nichž vede hrana na vrchol x, W je množina vrcholů na které vedou hrany z vrcholu x a p představuje míru, se kterou autority ovlivňují jiné autority ($0 \leq p \leq 1$).

$$\begin{aligned} BHA_x &= (1-p) \cdot \sum_{u \in U} BHH_u + p \cdot \sum_{u \in U} BHA_u \\ BHH_x &= \sum_{w \in W} BHA_w \end{aligned} \tag{27}$$

Balanced SALSA, či B-SALSA obsahuje stejnou úpravu jako B-HITS. Znázorněna je ve vzorci (28), kde BSA_x je B-SALSA autoritové skóre vrcholu x, BSH_x je B-SALSA rozcestníkové skóre vrcholu x, M_w je počet vstupních hran vrcholu w a ostatní proměnné jsou stejné jako v předchozích případech.

$$\begin{aligned} BSA_x &= (1-p) \cdot \sum_{u \in U} \frac{BSH_u}{N_u} + p \cdot \sum_{u \in U} \frac{BSA_u}{N_u} \\ BSH_x &= \sum_{w \in W} \frac{BSA_w}{M_w} \end{aligned} \tag{28}$$

Dále upravují algoritmus SCEAS a zavádějí jeho varianty *SCEAS – Publication Score* (PS), který napravuje nevýhodu Prestiže, *SCEAS – Balanced Publication Score* (BPS), *SCEAS – Exponentially Weighted Publication Score* (EPS) a *SCEAS – Balanced Exponentially Weighted Publication Score* (BEPS).

Dle experimentu, který autoři prováděli, konvergovaly varianty SCEAS1 a SCEAS2 nejrychleji ze všech variant SCEAS a PageRank a B-HITS konvergovaly rychleji, než HITS, B-SALSA, Prestiž a SALSA. Dále ukazují, že SCEAS1, SCEAS2 a BESP mají téměř totožné výsledky (odchylka max. 0,4%). V porovnání s držiteli *ACM SIGMOD E.F. Codd Innovations Award* nejlepších výsledků dosáhly PageRank a BPS, nejhorších Prestiž (5x horší) a BHA (3,5x horší).

2.3.8 Centrality Measure

Centralita je, vedle popularity a prestiže, další mírou významnosti, kterou lze v grafu určovat. Pojem *centralita* pochází z oblasti sociologie a první zmínku o ní můžeme nalézt v (Freeman, 1977), kde autor definoval sadu metod či mír centrality (*Centrality Measures*) založených na betweenness. V následující práci (Freeman, 1979) již autor definoval *degree*, *closeness*, *betweenness* a *eigenvector centralitu*, které si přiblížíme více. Popis dalších mír centrality, které rozšiřují zmíněné základní míry centrality, lze nalézt např. v (Hanneman & Riddle, 2005), kde autoři také představují software UNICET[5] pro výpočet všech uvedených metrik. Odkazy na použití mír centrality pro vyhodnocení různých grafů lze nalézt v (Yan & Ding, 2009), kde autoři aplikují míry centrality i na graf spoluautorství.

Degree Centrality

Nejjednodušší mírou centrality je *Degree* („stupeň“) centralita (Freeman, 1979; Yan & Ding, 2009), která zastupuje počet hran nebo součet vah hran, které vrchol má s jinými vrcholy. Pokud je graf orientovaný, lze rozlišovat *in-degree* a *out-degree* centralitu, kde „in“ zastupuje vstupní hrany a „out“ hrany výstupní.

Obecně uvažujeme, že vrchol s vysokým počtem hran nebo více spojeními je ve struktuře grafu více centrální a má tak větší schopnost ovlivňovat ostatní. Vrchol (např. autor), na který vede mnoho hran (vysoké in-degree), lze označit za prominentní, přední, či populární vrchol. Vrchol, ze kterého vede mnoho hran (vysoké out-degree), lze naopak označit za vlivný vrchol – má vyšší šanci ovlivnit ostatní.

Pokud bychom chtěli porovnávat vrcholy různých grafů, museli bychom hodnoty degree centrality normalizovat. To provedeme tak, že vydělíme degree vrcholů maximálním možným počtem hran, které vrchol může mít, tj. *(n-1)*, kde *n* je počet všech vrcholů grafu (Ferrara, 2012; Freeman, 1979).

Vedle základní Freemanovy definice degree centrality existuje i definice Bonacichova, která využívá i vazby přátel (zmínka v (Hanneman & Riddle, 2005)).

Closeness centrality

Closeness („blízkost“) centralitu (Freeman, 1979; Yan & Ding, 2009) lze chápat jako míru toho, jak dlouho bude trvat, než se informace rozšíří z daného vrcholu do všech ostatních vrcholů grafu, či jak blízko je vrchol ke všem ostatním vrcholům grafu. Lze ji tedy počítat pouze pro souvislé komponenty. Pokud celý graf není jednou souvislou komponentou, lze vypočítat closeness centralitu ve všech jeho komponentách zvlášť a následně vypočítané hodnoty vrcholů normalizovat velikostí komponent, tj. vynásobit je *(n-1)*, kde *n* je počet vrcholů komponenty, ve které se vrchol nachází (Freeman, 1979). Tímto způsobem lze porovnávat i closeness centrality vrcholů různých grafů. Vzorcem closeness centrality je vzorec (29), kde $C_c(u)$ je skóre closeness centrality vrcholu *u, V* je množina všech vrcholů

[5] Knihovna *UNICET* - https://sites.google.com/site/ucinetsoftware/home

grafu a *d(u,v)* je délka nejkratší cesty z vrcholu *u* do vrcholu *v*. Čím blíže je vrchol všem ostatním vrcholům grafu, tím má vyšší hodnotu closeness centrality.

$$C_C(u) = \sum_{v \in V} \frac{1}{d(u,v)} \tag{29}$$

Pokud je graf vážený, musíme znát význam vah hran. Jestliže váhy hran vyjadřují vzdálenost (čím větší váhu hrana má, tím jsou od sebe vrcholy dále), tak výpočet neměníme. Pokud ale váhy hran vyjadřují spříznění či blízkost (čím větší váhu hrana má, tím blíže si jsou její koncové vrcholy, např. počet citací), tak ve výpočtu vzdálenosti *d(u,v)* musíme počítat s obrácenými hodnotami vah hran (tj. *1/d(u,v))* nebo jednodušeji umocnit $C_c(u)$ na mínus první.

Definice a vzorec closeness centrality byly také zkoumány a upravovány. Úpravy autorů Hubbell, Katz, Taylor, Stephenson a Zelenovu míru vlivu lze nalézt v (Hanneman & Riddle, 2005).

Betweenness centrality

Betweenness („mezilehlost“) centralita (Freeman, 1979; Yan & Ding, 2009) vyjadřuje schopnost vrcholu propojovat (rozdílné) skupiny vrcholů. Vrchol s vysokou betweenness centralitou má významnou roli v propojování odlišných skupin. Příkladem může být osoba, která navštěvuje dva zájmové kroužky (např. volejbal a fotbal) a tím propojuje dvě skupiny osob – pokud tato osoba bude jedinou osobou, která navštěvuje oba kroužky, tak v grafu známosti, tvořeném z osob obou kroužků, bude mít největší betweenness centralitu. Tato osoba může do značné míry ovlivňovat dění v grafu např. blokováním (nežádoucích) zpráv, vybíráním poplatků „za spojení“, či může izolovat některé osoby, které nemají jinou možnost jak se ke sdílené informaci dostat.

Při výpočtu betweenness centrality nás zajímá, na kolika nejkratších cestách mezi dvěma různými vrcholy (počítáno přes všechny možné dvojice vrcholů) vrchol leží. Výpočet znázorňuje vzorec (30), kde $C_B(u)$ je betweenness centralita vrcholu *u*, $g_{j,u,k}$ je počet nejkratších cest mezi vrcholy *j* a *k*, které vedou přes vrchol *u* a $g_{j,k}$ je počet všech nejkratších cest mezi vrcholy *j* a *k*.

$$C_B(u) = \sum_{j,k \in V; j,k \neq u} \frac{g_{j,u,k}}{g_{j,k}} \tag{30}$$

Porovnávat hodnoty betweenness centrality vrcholů různých grafů lze po jejich normalizaci. Tu provedeme vydělením hodnot betweenness centralit počtem všech možných hran, které by graf mohl obsahovat, tj. *(n-1)*(n-2)* pro orientovaný graf a *(n-1)*(n-2)/2* pro graf neorientovaný, kde *n* je počet vrcholů grafu (Freeman, 1979). Je-li graf vážený, tak při zjišťování nejkratších cest musíme opět brát v úvahu význam hran a správně určit nejkratší cesty, jak již bylo zmíněno u closeness centrality.

Další varianta betweennes centrality využívá všech cest mezi dvěma vrcholy, ne pouze těch nejkratších, a tím uvažuje, že při komunikaci mezi dvěma vrcholy nemusí být využito pouze nejkratších cest (tj. je-li nejkratší cesta nedostupná, použije se druhá nejkratší, pokud existuje, atd.) (Hanneman & Riddle, 2005). Tento přístup je ale výpočetně náročný.

Eigenvector centrality

Při výpočtu *Eigenvector* („vlastní vektor“) centrality (Bonacich, 1972; Newman, 2010; Ruhnau, 2000) uvažujeme, že významný vrchol má významné sousední vrcholy. Ve své první variantě byla eigenvector centralita počítána dle vzorce (31), kde $\boldsymbol{A}$ je řádkově normalizovaná matice sousednosti, $\boldsymbol{x}$ je vlastní vektor matice $\boldsymbol{A}$ a λ je vlastní číslo odpovídající řešení. Rovnice má více řešení, ale nás zajímá pouze to řešení, jehož všechny složky jsou nezáporné. Dle Perron-Frobeniovy věty (Langville & Meyer, 2006) pro každou nezápornou primitivní matici existuje právě jedno řešení, ve kterém vektor $\boldsymbol{x}$ obsahuje pouze nezáporné hodnoty, a to odpovídá největšímu vlastnímu číslu.

$$\boldsymbol{Ax} = \lambda \boldsymbol{x} \tag{31}$$

Vzorec (31) lze rozepsat do rekurentního tvaru pro jednotlivé složky, jak ukazuje vzorec (32), kde $x(u)_i$ je hodnota eigenvector centrality vrcholu u v iteraci i.

$$x(u)_{i+1} = \frac{1}{\lambda} \sum_{v \in V} A_{uv} \cdot x(v)_i \tag{32}$$

Protože definice eigenvector centrality velmi blízce odpovídá definici PageRanku (viz část 2.2), kde také uvažujeme význam odkazujících vrcholů, je eigenvector centralita PageRankem často nahrazována (Kurtz & Bollen, 2011; West, 2010), protože PageRank navíc řeší některé další problémy.

2.3.9 Další PageRanku podobné algoritmy

Další algoritmy, které se svou podstatou podobají algoritmu PageRank lze nalézt např. v (Borodin et al., 2005), kde autoři upravují algoritmus HITS a ukazují jeho nové varianty zmírňující některé jeho nedostatky. V (Yan & Ding, 2010) autoři určují prestiž publikací na základě Impact Factoru časopisů a času vzniku citace (publikování). Tento postup následně rozšiřují v článku (Yan et al., 2010), kde ukazují algoritmus *P-Rank*, s jehož využitím hodnotí současně publikace, autory a časopisy. Obdobný postup hodnocení více bibliografických entit současně využitím provázaných vzorců lze nalézt také v (Yu et al., 2012), ovšem tento postup je aktuálně pouze teoretický (k hodnocení chybí některá data jako např. komentáře uživatelů). Posledním algoritmem, o kterém se v této souvislosti zmíníme, je algoritmus *CoRank* (Zhou et al., 2007), který hodnotí současně autory a publikace.

Algoritmus *CiteRank* (Walker et al., 2006) je další úpravou PageRanku pro hodnocení publikací na základě jejich citačního grafu. Úpravou autoři uvažují skutečnost, že nová citace (znamenající, že článek je významný v aktuální linii výzkumu) má větší váhu, než citace stará, a proto exponenciálně snižují váhy citací v závislosti na jejich stáří. Pro vyhledávání expertů v online znalostní komunitě autoři (G. A. Wang et al., 2013) navrhli algoritmus *ExpertRank*, který určuje relevanci autorů na základě podobnosti obsahů jejich publikací s uživatelovým dotazem.

2.4 Možnosti přerozdělování významnosti

Občas nás zajímá, jak lze přenést hodnocení z jednoho typu entit na jiný typ entit, který s prvním typem souvisí (dále bude použito označení „přerozdělování významnosti“), např. z publikací na autory, z publikací na časopisy, z autorů na publikace, či instituce atd.

V (Assimakis & Adam, 2010) je zmíněno několik možností přerozdělování významnosti publikací jejich autorům. Mezi základní možnosti patří:

a) Sdílení hodnoty publikace všemi spoluautory, tj. každý autor dostane celou hodnotu své publikace, bez ohledu na počet spoluautorů.
b) *Rovnoměrné* (či *uniformní*) *rozdělení* hodnoty publikace mezi její autory, tj. pokud má publikace *n* autorů, tak každý získá *1/n* z její hodnoty.
c) *Nerovnoměrné rozdělení* hodnoty publikace mezi její autory, tj. obvykle první autor získá větší díl z hodnoty publikace, než další její autoři.

Hodnoty, které autor získá ze svých publikací, se pro daného autora sčítají. Dále se budeme věnovat variantě c), kde bylo navrženo lineární, geometrické a zlaté rozdělení (Assimakis & Adam, 2010).

Lineární rozdělení znázorňuje vzorec (33), kde $lr_{n,j}$ je lineární rozdělení pro *j*-tého autora z *n* autorů publikace. Př.: pokud publikace má 3 autory, tak první autor získá 50%, druhý 33% a třetí 17% z její hodnoty.

$$lr_{n,j} = \frac{-2}{n \cdot (n+1)} j + \frac{2}{n} \tag{33}$$

Geometrické rozdělení využívá vzorců (34), kde *n* je počet autorů publikace, *j* pozice daného autora a λ musí být reálné kladné číslo (dle důkazu v (Assimakis & Adam, 2010) existuje právě jedno řešení, které je reálné a kladné). Př.: pokud publikace má 3 autory, tak λ=0,5437 a první autor získá 54%, druhý 30% a třetí 16% z hodnoty publikace.

$$\begin{gathered} \lambda^n + \lambda^{n-1} + \cdots + \lambda^2 + \lambda = 1 \\ gr_j = \lambda^j \end{gathered} \tag{34}$$

Zlaté rozdělení využívá vzorce (35), kde φ=0,618 (získáno z $\varphi^2 + \varphi = 1$) je „zlatý index“. Př.: pokud publikace má 3 autory, tak první autor získá 62%, druhý 24% a třetí 15% z její hodnoty. Pokud má publikace 4 autory, tak první a druhý autor získají stejná procenta jako v předchozím případě, třetí autor získá 9% a čtvrtý autor 6% z hodnoty publikace.

$$zr_{n,j} = \begin{cases} 1, & n = 1 \\ \varphi^{2 \cdot j - 1}, & j = 1, \dots, n-1, \quad n \geq 2 \\ \varphi^{2 \cdot j - 2}, & j = n, \quad n \geq 2 \end{cases} \tag{35}$$

Autoři (Assimakis & Adam, 2010) využívají zlatý index k určení produktivity autorů publikací. Každá publikace má hodnotu jedna a součet zlatých indexů daného autora autoři nazývají *p-index*.

2.5 Bibliografické databáze, vytvářené grafy a možnosti porovnání výsledných pořadí

Cílem této kapitoly je ukázat, které informace z bibliografických databází můžeme použít pro tvorbu grafu. Vyhodnocovaný typ grafu udává vlastnost, která je hodnotící metrikou měřena. Dále zmíníme několik bibliografických databází a možnosti porovnání výsledných pořadí.

2.5.1 Bibliografické grafy

Bibliografickým grafem rozumíme graf vytvořený z bibliografických záznamů, tj. vrcholy představují zvolené entity (publikace, autory, instituce atd.) a hrany jejich vzájemnou interakci. Vyhodnocování můžeme rozdělit na vyhodnocení „homogenního" grafu (tj. všechny vrcholy a hrany jsou pouze jednoho typu), multi-dimensionální vyhodnocení (Sayyadi & Getoor, 2009; Yu et al., 2012) (tj. vyhodnocení, které pracuje s více typy grafů) a vyhodnocení multi-typového (Yang et al., 2010) či heterogenního (Yan et al., 2010) grafu (tj. grafu, který obsahuje vrcholy i hrany různého typu). V některých případech může vyhodnocovaný homogenní graf vzniknout kombinací více homogenních grafů a určení, o který typ vyhodnocení se jedná, není jednoznačné, což ale obvykle není příliš důležité.

Základními entitami jsou publikace (článek, kniha, referát atd.) a základní interakcí jejich vzájemné citace. Z těchto dat lze vytvořit *citační graf publikací*, kde hrana vždy vede od citující publikace k citované. Použití citačního grafu publikací ukazuje např. (Ma et al., 2008). Každá publikace obvykle obsahuje záznam o svých autorech, jejich afiliacích a časopisu nebo konferenci, kde byla publikována, popř. místu publikování. Ze všech těchto entit lze vytvořit citační grafy (např. pokud publikace autora A cituje publikaci autora B, tak lze vytvořit *citační graf autorů* s vrcholy A a B, kde vede hrana z vrcholu A do vrcholu B). Z afiliací autorů lze často zjistit příslušný stát a vytvořit tak *citační graf států*. Z názvu publikace lze odvodit příslušné téma publikace a vytvořit tak *citační graf témat*. Stejným způsobem lze použít i klíčová slova publikací. Citační grafy časopisů, institucí a států byly použity v (Egghe, 2010; Leydesdorff, 2013), citační grafy autorů a institucí např. v (Nykl, 2011), citační graf států v (Fiala, 2012a). Důležitým faktorem ovlivňujícím tvorbu citačních grafů autorů a odvozených entit je, zda použijeme vždy pouze autora, který byl v publikaci uveden na první pozici, nebo použijeme všechny uvedené autory (Zhao, 2005).

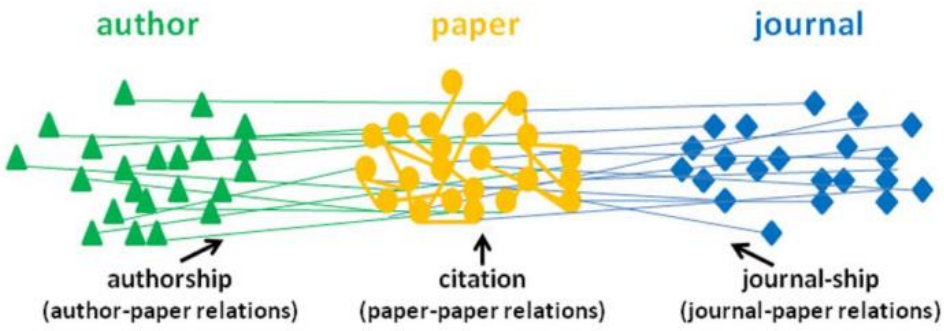
Vedle citačních grafů lze z bibliografických záznamů vytvářet *grafy spoluautorství, spolupráce* či *společného výskytu*, kde mezi entitami vede neorientovaná hrana, pokud se nachází ve stejném záznamu o publikaci. Tímto způsobem lze vytvářet grafy spolupráce autorů, institucí nebo států a grafy společného výskytu témat, klíčových slov či slov obsažených v názvu publikace. Vyhodnocení grafu spolupráce autorů ukazuje např. (Yan & Ding, 2009).

Dalšími vytvářenými grafy jsou *grafy společně citovaných* (*co-citation* nebo *co-cited*) a *společně citujících* (*co-citing* nebo *co-reference*) entit. V grafu společně citovaných vede mezi dvěma entitami neorientovaná hrana, pokud obě byly citovány ve stejné publikaci. V grafu společně citujících vede mezi dvěma entitami neorientovaná hrana, pokud obě citují stejnou publikaci. Vyhodnocení grafu společně citovaných autorů nalezneme např. v (Y. Ding & Yan, 2009; Yan & Ding, 2012). Některé další vytvářené grafy můžeme nalézt v (Yan & Ding, 2012), kde autoři tyto grafy vzájemně porovnávají.

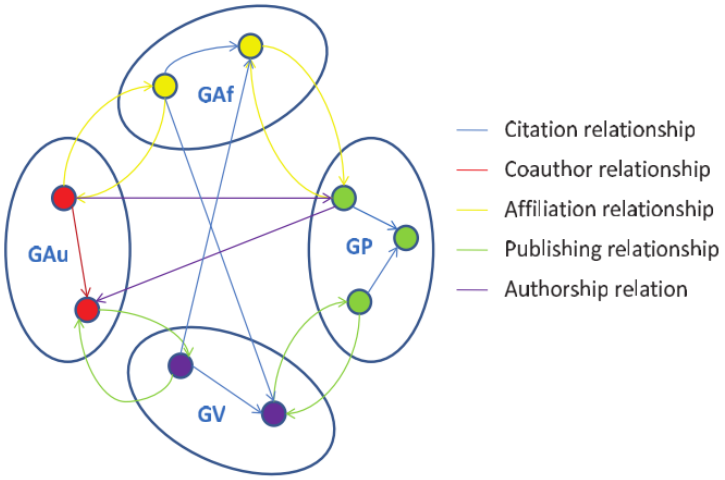
Za multi-dimensionální vyhodnocení můžeme označit takové vyhodnocení, které pracuje s více typy grafů. V (Sayyadi & Getoor, 2009) autoři s využitím PageRanku vyhodnocují citační graf publikací a následně aplikují algoritmus HITS na bipartitní graf autorství (tj. autoři a jejich publikace), aby získali hodnocení autorů i publikací. V (Yu et al., 2012) autoři současně (s využitím soustavy rovnic) hodnotí publikace, autory, komentáře a zdroje, tj. časopisy či konference.

Vyhodnocování multi-typového či heterogenního grafu je spíše idea a většinou se jedná o multi-dimensionální vyhodnocení. Graf je multi-typový či heterogenní pokud obsahuje více typů vrcholů nebo více typů hran, či oboje. Tuto vlastnost autoři obvykle ve svých pracích nastíní a ukáží

vytvořený graf, ale poté tento graf vyhodnocují po částech, tj. vyhodnocují vlastně několik grafů, stejně jako u multi-dimensionálního vyhodnocení. Částečnou výjimku tvoří bipartitní grafy, kde ale hrany nikdy nevedou mezi vrcholy stejné množiny. Vyhodnocení heterogenního grafu ukazuje např. (Yan et al., 2010), kde autoři používají graf (na obr. 3) složený z citačního grafu publikací, bipartitního grafu autorství a bipartitního grafu publikování (tj. časopisy a publikace v nich obsažené). Takto vytvořený graf nazývají heterogenním grafem, ale následně ho vyhodnocují po částech s využitím právě zmíněných tří grafů. Stejným postupem pracují autoři (Yang et al., 2010), kteří ukazují heterogenní graf (na obr. 4) vytvořený spojením citačního grafu autorů a grafu spoluautorů (G_{Au}), citačního grafu publikací (G_P), citačního grafu institucí (afiliací) a grafu spolupráce institucí (G_{Af}), citačního grafu míst publikování (G_V), bipartitního grafu publikace-autoři (autorství), bipartitního grafu autoři-afiliace (příslušnost), bipartitního grafu autoři-místa publikování, bipartitního grafu publikace-afiliace, bipartitního grafu publikace-místa publikování a bipartitního grafu afiliace-místa publikování. Protože obě zmíněné práce následně vyhodnocují dílčí podgrafy, můžeme tyto přístupy označit také za multi-dimensionální vyhodnocení.



Obr. 3: Heterogenní graf - citace publikací, autorství a publikování. Přejato z (Yan et al., 2010).



Obr. 4: Heterogenní graf přejatý z (Yang et al., 2010).

Některá vyhodnocení pracují s homogenním grafem, který vznikl spojením více typů grafu, obvykle dvou. Například v článku (Fiala et al., 2008) autoři používají citační graf publikací a bipartitní graf autorství, s jejichž využitím vytváří citační graf autorů a ten vyhodnocují.

2.5.2 Bibliografické databáze

Nejznámějšími bibliografickými databázemi (dále jen databáze) jsou Web of Science, Scopus, Google Scholar, CiteSeer, DBLP, Microsoft Academic Search a arXiv. Následující informace o databázích jsou čerpány z oficiálních webů databází, jejich vyhledávání a z (Bar-Ilan, 2007; Bellis, 2009; Fiala, 2011).

Web of Science[6] (WoS) multioborová databáze Ústavu pro vědecké informace (*Institute for Scientific Information* – ISI) udržovaná firmou *Thomson Reuters* je jednou z nejstarších databází. Aktuálně shromažďuje vědecké články z více než 12000 vlivných časopisů a více než 150000 sborníků konferencí a zastřešuje tak zhruba 250 vědních disciplín. Všechny vkládané časopisy a sborníky podléhají přijímacímu řízení. Indexovány jsou publikace od roku 1945. WoS byl mnohokrát použit v citační analýze, jak ukazují např. (Y. Ding, 2011; Fiala, 2012b; Yan & Ding, 2009, 2012, 2011; Zhu & Guan, 2013).

Scopus[7] vnikl v roce 2004 a je udržován firmou *Elsevier*. Jedná se o multioborovou databázi obsahující více než 50 miliónů záznamů o vědeckých publikacích z více než 21000 titulů (časopisy a sborníky) od zhruba 5000 vydavatelů. Indexovány jsou manuálně vložené publikace od roku 1960 ze všech vědních oborů. Scopus pro citační analýzu používají např. (Elkins et al., 2010; Franceschini et al., 2013; Haddow & Genoni, 2010).

Google Scholar[8] (GS) společnosti *Google Inc.* vznikl v roce 2004 a jedná se o automatický systém shromažďování informací o vědeckých článcích. Indexují se články vydavatelů ze všech vědních oborů, kteří poskytují alespoň abstrakt článků zdarma. Počet indexovaných článků ani rozsah let není znám. Přístup do vyhledávání je zdarma. Použit v citační analýze byl např. v (Amara & Landry, 2012; Bar-Ilan, 2007; Harzing, 2013; Mingers & Lipitakis, 2010).

CiteSeer[9], či jeho „nová generace“ *CiteSeer^X*, byl prvním autonomním systémem, který indexuje vědecké publikace v elektronické podobě (Giles et al., 1998). Vytvořen byl v *NEC Research Institute* (Princeton, NJ, USA). Dle (Fiala, 2011) je CiteSeer zaměřen na oblast počítačových věd a v roce 2010 obsahoval téměř 33 miliónů záznamů. Rozsah indexovaných let není z vyhledávání jednoznačně patrný, protože se zde projevují nedůslednosti v datech – některé články obsahují místo čtyřciferného údaje o roku publikování údaj pouze dvouciferný. CiteSeer^X je stále ve verzi beta. Přístup do vyhledávání je zdarma. Použití CiteSeeru v citační analýze nalezneme např. v (Fiala, 2011, 2012a; Nykl & Ježek, 2012; Nykl, 2011; Sidiropoulos & Manolopoulos, 2005b; Zhou et al., 2007).

DBLP[10] (Data Bases and Logic Programming) databáze university v Trieru (Německo) vznikla v roce 2005. Dříve zahrnovala, jak název napovídá, oblast databázových systémů a logického programování, ale dnes se soustředí na celou oblast počítačových věd. Databázi lze stáhnout v podobě XML souborů. Aktuálně obsahuje téměř 2,4 milionu manuálně vložených záznamů od roku 1936 [11]. Přístup do vyhledávání je zdarma. DBLP byla mnohokrát použita v citační analýze, např. (Di Caro et al., 2012; Fiala et al., 2008; Liu et al., 2005; Nykl & Ježek, 2012; Sidiropoulos & Manolopoulos, 2005a, 2006).

[6] Databáze *Web of Science* - http://www.webofknowledge.com
[7] Databáze *Scopus* - http://www.scopus.com
[8] Databáze *Google Scholar* - http://scholar.google.com
[9] Databáze *CiteSeer* - http://www.citeseer.com
[10] Databáze *DBLP* - http://dblp.uni-trier.de
[11] Statistiky vztahující se k databázi DBLP - http://dblps.uni-trier.de/~mwagner/statistics/

Microsoft Academic Search[12] (MAS) společnosti *Microsoft* vnikl v roce 2009 a obsahuje více než 48 miliónů publikací od více než 20 miliónů autorů ze 14 oblastí výzkumu. Lze v něm nalézt i články Isaaca Newtona z roku 1687. Indexace publikací je automatická a přístup do vyhledávání je zdarma. Použití MAS v citační analýze lze nalézt v (Jacsó, 2011).

arXiv[13] vznikl v roce 1991 pod záštitou knihovny Cornellovy univerzity (Ithaca, NY, USA) jako automatizovaný elektronický archív a distribuující server vědeckých článků. Zahrnuje 6 oblastí výzkumu (fyzika, matematika, počítačové vědy, nelineární vědy, kvantitativní biologie a statistika) a obsahuje články od roku 1992. Počet indexovaných článků není uveden [14]. Přístup do vyhledávání je zdarma. Použití arXiv v citační analýze lze nalézt v (Sayyadi & Getoor, 2009).

Vedle výše zmíněného základního porovnání těchto databází, lze nalézt i jejich porovnání při použití v citační analýze. Autoři (Mingers & Lipitakis, 2010) porovnávají WoS a GS v oblasti byznysu a managementu a dochází k závěru, že GS zastřešuje tuto oblast více než WoS. Autor (Harzing, 2013) porovnává WoS a GS z pohledu vědních oborů využitím držitelů Nobelovy ceny a dochází k závěru, že GS je méně zaujatý než WoS a může tak napravit znevýhodněné postavení sociálních věd v citační analýze. Autorka (Bar-Ilan, 2007) porovnává výpočet h-indexu Izraelských vědců na základě dat získaných z WoS, Scopus a GS, ale její závěr není jednoznačný.

Za zmínku stojí, že dle nařízení České vlády pro roky 2010 až 2012 (Úřad vlády ČR, 2012) se pro hodnocení výzkumných organizací v části publikační činnosti používají publikace zaznamenané v RIV (Rejstřík informací o výsledcích), které se nachází v databázích WoS, Scopus nebo ERIH (humanitní obory), či se nachází na seznamu Českých recenzovaných neimpaktovaných periodik. Při rozdělování bodů za publikace se u časopisů zařazených do WoS uvažuje Impact Factor a u časopisů obsažených pouze v jiných databázích se uvažuje daná databáze (např. za článek v časopise indexovaném ve Scopus autor získá vždy 12 bodů). Tento systém se ale pro roky 2013 až 2015 mění (Úřad vlády ČR, 2013).

2.5.3 Možnosti porovnání výsledných pořadí

Pokud jsme již vytvořili požadovaný graf a vyhodnotili ho zvolenými metrikami, zajímá nás, jak lze porovnat výsledná pořadí. Obvyklým cílem je buďto určení, která z použitých metrik poskytuje lepší výsledné pořadí, nebo pouhé porovnání podobnosti výsledků. V následujícím textu této části pro názornost uvažujme, není-li řečeno jinak, porovnání dvou výsledných pořadí autorů vědeckých publikací, která vynikla vyhodnocením citačního grafu autorů.

Nejjednodušší metodou porovnání výsledných pořadí je porovnání několika nejlepších pozic, např. prvních dvacet pozic výsledných pořadí. Zde se ptáme, kolik autorů je na nejlepších pozicích v obou výsledných pořadích a jak se liší jejich umístění. Toto vyhodnocení můžeme nalézt např. v (Yan & Ding, 2009).

Další možností porovnání výsledných pořadí je statistické porovnání jejich podobnosti. K tomuto účelu lze použít koeficienty korelace, přesněji Spearmanův (Spearman, 1904) nebo Kendallův (Kendall, 1938) koeficient pořadové korelace, které měří statistickou závislost dvou veličin. Veličinou zde rozumíme posloupnost prvků s určeným pořadím, přičemž obě zkoumané veličiny musí

[12] Databáze *Microsoft Academic Search* - http://academic.research.microsoft.com
[13] Databáze *arXiv* - http://www.arxiv.org
[14] Statistiky vztahující se k databázi *arXiv* - http://arxiv.org/help/stats/2012_by_area/index

obsahovat totožné prvky. Porovnání je následně závislé pouze na vytvořeném pořadí a ne na hodnotách, dle kterých pořadí vzniklo. Výsledný koeficient pořadové korelace, který může nabývat hodnot z intervalu <-1;1>, udává, do jaké míry jsou na sobě veličiny funkčně závislé. Přesněji, korelační koeficient o hodnotě jedna říká, že obě veličiny jsou na sobě zcela funkčně závislé, nulová hodnota říká, že mezi zkoumanými veličinami není žádná funkční závislost, a hodnota mínus jedna značí opačnou funkční závislost (tj. prvek, který je v první veličině na první pozici, bude ve druhé veličině na pozici poslední atd.). Nejčastěji používaným koeficientem korelace pro porovnání výsledků citační analýzy je Spearmanův koeficient. Jeho použití nalezneme např. v (Y. Ding & Yan, 2009; Fiala et al., 2008; Ma et al., 2008).

Chceme-li určit, která metoda hodnocení poskytuje „lepší" výsledné pořadí, musíme nejprve určit, co znamená, že nějaké pořadí je lepší, než pořadí jiné. Musíme tedy zvolit referenční pořadí či hodnocení, které prohlásíme za nejlepší, a porovnávat, jak blízké je námi získané výsledné pořadí k tomuto referenčnímu pořadí. V oblasti hodnocení časopisů či institucí narazíme na problém, že žádné referenční hodnocení neexistuje, vyjma žebříčku univerzit[15] (který se ale každý rok mění). V oblasti hodnocení autorů lze jako referenční hodnocení použít různá ocenění udílená za vědeckou a publikační činnost, jako např. Nobelova cena udílená ve zkoumané oblasti (použití např. v (Harzing, 2013)). Pokud námi zkoumaná oblast výzkumu jsou počítačové vědy, můžeme použít Turingovu cenu (*ACM A.M. Turing Award*[16], použita např. v (Fiala, 2012b; Nykl & Ježek, 2012)), Coddovu cenu (*ACM SIGMOD E.F. Codd Innovations Award*[17], použita např. v (Fiala et al., 2008; Nykl, 2011; Sidiropoulos & Manolopoulos, 2005a)), cenu *VLDB 10 Year Award*[18] nebo *ACM Test of Time*[19] (užité např. v (Sidiropoulos & Manolopoulos, 2005a, 2006)), které obě mohou být použity i pro porovnání výsledných pořadí publikací. Hodnotící metrikou, dle které se následně porovnávají výsledná pořadí, může být součet, průměr, minimum, či maximum z pozic, které ve vypočítaném pořadí obsadili držitelé zvoleného typu ocenění.

2.6 Možnosti porovnání grafů

Několik metrik umožňujících porovnání dvou a více grafů či sítí z pohledu jejich struktury lze nalézt v (Ferrara, 2012; Hanneman & Riddle, 2005).

Poloměr grafu je dán délkou nejdelší z nejkratších cest grafu. *Průměrný stupeň vrcholu* je podílem počtu hran a počtu vrcholů grafu. *Hustota grafu* je podílem počtu hran grafu a počtu všech možných hran, které by graf mohl obsahovat, tj. *n*(n-1)* orientovaných hran, kde *n* je počet vrcholů grafu. Pokud je graf neorientovaný, je počet všech možných hran poloviční.

Rozložení stupňů vrcholů grafu se znázorňuje diagramem, kde na vodorovné ose je vynesen stupeň vrcholu a na svislé ose počet vrcholů, které mají daný stupeň. Pokud se jedná o klesající funkci (tj. se vzrůstajícím stupněm vrcholu klesá počet vrcholů s daným stupněm) řídící se mocninným zákonem (Barabási, 2005), lze míru poklesu zapsat jako $N(k)=k^{-\gamma}$, kde *N(k)* je počet vrcholů majících stupeň *k*, *k* je stupeň vrcholu a γ je exponent konektivity. U komplexních sítí či grafů platí mocninný zákon a nás obvykle zajímá právě exponent konektivity (Barabási, 2005).

[15] Web s hodnocením univerzit celého světa - http://www.webometrics.info
[16] Web *ACM A. M. Turing Award* - http://amturing.acm.org
[17] Web *ACM SIGMOD Edgar F. Codd Innovations Award* - http://www.sigmod.org/sigmod-awards
[18] Web *VLDB 10 Year Award*, např. - http://www-nishio.ist.osaka-u.ac.jp/vldb/archives/public/10year/10year.html
[19] Web *ACM Test of Time* - http://www.sigmod.org/sigmod-awards/sigmod-awards#time

Dalšími metrikami, které se používají k porovnání grafů, jsou např. *koeficient shlukování* a *zastoupení trojúhelníků* (podíl počtu trojúhelníků v grafu a všech možných trojúhelníků v daném grafu).

Pro výpočet těchto metrik a obvykle spousty dalších, např. mír centrality, lze použít exitující programové knihovny, jako např. *UNICET*[20], *Pajek*[21] nebo *Gephi*[22], které všechny mají rozsáhlou dokumentaci. Pajek je navíc použitelný pro analýzu velmi objemných grafů a Gephi pro vizualizaci grafů.

2.7 Použití PageRanku v různých oblastech výzkumu

Od svého vzniku byl algoritmus PageRank, či některá jeho upravená varianta, aplikován na grafy získané z různých oblastí výzkumu. Za zmínku stojí použití PageRanku v textových úlohách, kde vznikly např. *TextRank* (Mihalcea & Tarau, 2004), *LexRank* (Erkan & Radev, 2004) a *MFSRank* (López et al., 2011). Dále jeho uplatnění v klasifikaci (Dostal et al., 2014; Nykl et al., 2013), shlukování (Avrachenkov et al., 2008), štítkování či extrakci klíčových slov nebo frází z textu (Ortiz & Pinto, 2010) a v oblasti odstraňování dvojsmyslů jmen (*name disambiguation*) (Smirnova et al., 2010).

V úloze hodnocení genů vznikl *GeneRank* (Benzi & Kuhlemann, 2012; Morrison et al., 2005) a graf interakce proteinů byl vyhodnocován algoritmem *PageRank Affinity* (Voevodski et al., 2009). Vyhodnocení reputace v P2P sítích s využitím PageRanku ukazuje (Chirita et al., 2004). Použití PageRanku pro predikci další hrany v sociální síti můžeme nalézt v (Liben-Nowell & Kleinberg, 2007), určování reputace v sociální síti v (Han et al., 2012; Hao et al., 2012), hlasovací systém využívající sociální síť a PageRank v (Boldi et al., 2009) a vyhodnocení firemní e-mailové komunikace v (Berchenko et al., 2011).

Uplatnění PageRanku ve webovém vyhledávání a bibliometrii jsme vynechali, protože bylo zmíněno již dříve.

[20] Knihovna *UNICET* - https://sites.google.com/site/ucinetsoftware/home
[21] Knihovna *Pajek* - http://pajek.imfm.si/doku.php
[22] Knihovna *Gephi* - https://gephi.org

3 Vlastní dosažené výsledky

PageRank a hodnocení autorů vědeckých publikací

Hodnocením autorů vědeckých publikací jsem se zabýval už v diplomové práci (Nykl, 2011), kde jsem z dat DBLP (2004) a CiteSeeru (2005) vytvářel citační grafy autorů a afiliací, graf spoluautorství a graf spolupráce afiliací, které jsem následně vyhodnocoval PageRankem. Testovány byly různé hodnoty faktoru tlumení, ukončovacího kritéria výpočtu a ošetření slepých vrcholů z pohledu:

- počtu iterací algoritmu potřebných k dosažení požadované přesnosti výsledku
- vzájemné podobnosti výsledných pořadí

Testovány byly také PageRank s personalizací zastoupenou počtem vstupních hran vrcholů a PageRank uvažující váhy hrany. Všechna výsledná pořadí autorů byla vzájemně porovnána využitím Spearmanova a Kendallova koeficientu korelace a držitelů *ACM SIGMOD E.F. Codd Innovations Award*. Ukázalo se, že se vzrůstajícím faktorem tlumení a požadovanou přesností výsledku roste i počet iterací, přičemž výpočet ošetřující slepé vrcholy vykoná obvykle o několik málo iterací více. Dále, že faktor tlumení 0,95 je ve většině případů nejvyšší použitelný, protože při vyšším faktoru tlumení rychle roste počet iterací vedoucích k dosažení požadované přesnosti výsledku. Totéž lze říci i o vzrůstající požadované přesnosti, kde se jako ideální jeví hodnota 10^{-5}, která je ale závislá na velikosti grafu. Porovnání výsledných pořadí autorů s držiteli *ACM SIGMOD E.F. Codd Innovations Award* nebylo, dle zpětného pohledu, provedeno zcela správně, proto zde nebude zmíněno.

V článku (Nykl & Ježek, 2012) jsme se také zabývali hodnocením autorů a zajímalo nás, jak se změní jejich výsledné pořadí, pokud je vytvořeno z hodnot PageRaku získaných z citačního grafu autorů nebo z hodnot, které autoři získali ze svých publikací (tj. nejprve se PageRankem vyhodnotil citační graf publikací a následně se hodnoty publikací převedly na autory). Cílem bylo zjištění, které vyhodnocení poskytuje nejlepší pořadí v porovnání s *ACM A.M. Turing Award*, *ACM SIGMOD E.F. Codd Innovations Award*, *ACM Fellows* a *ISI Highly Cited*. V citačním grafu publikací byly testovány varianty s a bez samocitací (tj. odstranily se hrany mezi publikacemi s alespoň jedním stejným autorem) a hodnoty PageRanku publikací byly mezi autory rozděleny tak, že autor získal buď součet PageRanků všech svých publikací, nebo se hodnoty publikací rovnoměrně rozdělily mezi jejich autory. V citačním grafu autorů byly testovány tři přístupy k samocitacím:

- všechny samocitace ponechány
- odstraněny samocitace v citačním grafu autorů
- odstraněny samocitace na úrovni publikací – citace mezi publikacemi s alespoň jedním stejným autorem byly smazány.

Také byly vyzkoušeny tři varianty vážených grafů autorů (pozn.: mezi dvěma autory mohou vést nanejvýš dvě orientované hrany):

- nevážený graf (tj. všechny hrany mají váhu jedna)
- váha hrany je tvořena součtem dílčích částí citací (tj. pokud publikace P1 autora A a publikace P2 autorů A a B citovaly publikaci autora C, tak z A povede na C hrana s váhou 1,5 a z B na C s váhou 0,5)

- váhy hran vyjadřují počet citací (tj. pokud tři publikace, kde jedním z autorů byl autor A, citovaly publikaci autora C, tak z A na C povede hrana s váhou 3)

Citační grafy autorů byly navíc vyhodnoceny PageRankem s personalizací danou počtem publikací autora a citační grafy publikací byly vyhodnoceny PageRankem s personalizací danou počtem autorů publikace. Data byla opět z DBLP (2004) a CiteSeeru (2005).

Ve vyhodnocení CiteSeeru bylo nejlepší pořadí autorů získáno aplikováním PageRanku bez personalizace na nevážený citační graf autorů, kde samocitace byly odstraněny na úrovni publikací. Vyhodnocení DBLP nebylo jednoznačné, ale v porovnání s *ACM Fellows* a *ISI Highly Cited* poskytoval nejlepší pořadí autorů PageRank s personalizací aplikovaný na stejný graf jako u CiteSeeru.

V krátkém článku (Heller et al., 2011) jsme zkoumali vliv penalizování cyklů délky dva a tři v citačním grafu autorů vyhodnocovaném PageRankem. Závěrem bylo, že se vzrůstající penalizací vah hran, které se účastní cyklu, vzrůstá rozdíl mezi výslednými pořadími autorů. Významná ocenění, zmíněná v předchozích odstavcích a v (Nykl & Ježek, 2012), k porovnání výsledných pořadí autorů použita nebyla, proto nelze říci, jaká míra penalizace poskytovala nejlepší výsledné pořadí autorů.

Souhrnně lze říci, že pro objektivní hodnocení autorů na základě dat z DBLP nebo CiteSeeru můžeme využít algoritmus PageRank s faktorem tlumení do 0,95, kde ukončovacím kritériem výpočtu je požadovaná přesnost výsledků 10^{-5}. V DBLP a CiteSeeru byly nejlepší výsledky získány z neváženého citačního grafu autorů, kde samocitace byly odstraněny na úrovni publikací. Tyto výsledky jsou však závislé na daných dvou databázích, a protože např. DBLP obsahuje velmi málo citací (což je zřejmě příčinou toho, že lepší výsledky zde poskytuje PageRank s personalizací), je v budoucnu důležité ověřit tyto postupy na databázi, která bude manuálně udržovaná a bude obsahovat velké procentuální zastoupení citací, jako např. Web of Science. Penalizace cyklů v grafu nadále používána nebude, protože malá penalizace nezpůsobuje příliš změn ve výsledném pořadí autorů a velká penalizace odstraňuje přirozené vlastnosti citování (tj. že prestižní autoři citují navzájem své publikace, protože jsou obvykle přínosem k poznání v dané oblasti výzkumu).

PageRank a volba vlastností využitím Linked Data

V oblasti volby vlastností jsme aplikovali PageRank na graf tvořený klíčovými slovy dokumentu s cílem nalezení vlastnosti, která nejlépe vystihuje jeho obsah. Vazby mezi termíny a další doplňující termíny byly získány z Linked Data (Lee, 2006), která navrhl Tim-Berners Lee jako prostředek pro konstrukci sémantického webu.

Našimi úkoly bylo:

1) navrhnout postup, s jehož využitím dokážeme množinu původních klíčových slov dokumentu rozšířit o další obecnější termíny, které s původními souvisí, aniž by tyto termíny byly v dokumentu použity.
2) určit významnost jednotlivých termínů zastupujících dokument a vybrat termín (nebo termíny), který bude nejvíce vystihovat téma dokumentu.

První úkol byl řešen dvěma způsoby. Nejprve se vytvořil graf z původních klíčových slov dokumentu - vazby mezi termíny byly nalezeny v Linked Data, přičemž byly využívány pouze vazby představující zobecnění, tj. vedoucí od specializovaných termínů k termínům obecnějším, např.

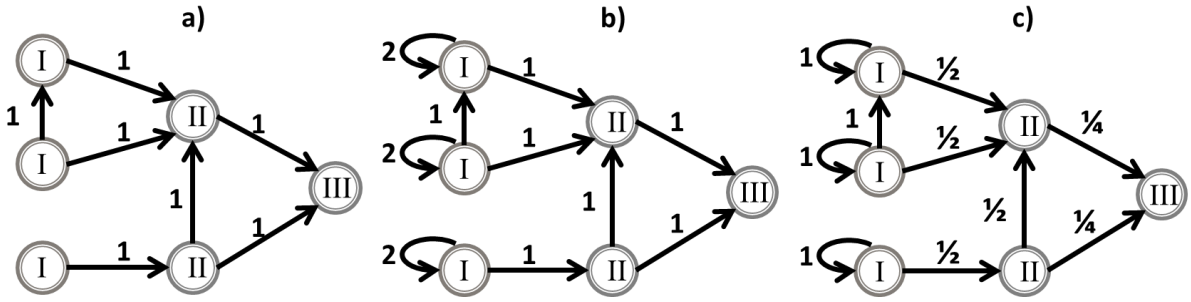
z *MySQL* na *Relační databáze*. Do grafu byly přidány nové termíny z Linked Data, na které z původních termínů vedly hrany. Problémem bylo určení, kdy se stávající graf už nebude rozšiřovat o další termíny. V prvním způsobu jsme rozšiřování grafu ukončili, pokud celý graf tvořil jednu souvislou komponentu. To se ukázalo dobře použitelné pro kolekci *Call for Papers* (ze které byly odstraněny multi-oborové konference), protože rozšiřováním původních klíčových slov dokumentu rychle vznikl souvislý graf (v téměř nejhorším případě např. různé pojmy z počítačových věd brzo odkázaly na vrchol *Počítačové vědy* a tím vznikla jedna souvislá komponenta).

Pro kolekci novinových článků nebyl tento postup příliš použitelný, protože např. kriminální článek obsahoval informace o místě zločinu a typu zločinu, které jsou z různých částí použitých Linked Data (např. pojem *gymnázium+jméno* vedl k doplnění pojmů *vzdělávací institut* a *„jméno města, kde se nachází*“, kdežto pojem *znásilnění* vedl k doplnění pojmu *sexuální zločin* – pokud bychom graf rozšiřovali, dokud nebude tvořit jednu souvislou komponentu, zahrnuli bychom do něj spoustu pojmů, které už nás nezajímají, a následně určený nejvýznamnější pojem by byl buďto velmi obecný, nebo zcestný). Proto jsme použili druhý způsob a to rozšiřování grafu pouze v případě, že libovolný pojem, z pojmů, které byly do grafu doplněny jako poslední, obdržel nejvyšší hodnotu PageRanku.

Druhý úkol spočíval v určení nejvýznamnějšího termínu ve vytvořeném grafu. Testovány byly tři varianty grafů, jak ukazují příklady na obr. 3:

a) nevážený (neupravený) graf – levý graf na obr. 3
b) graf, ve kterém původní klíčová slova dostala smyčku s dvojnásobnou váhou, než mají ostatní hrany grafu, aby byla tato slova ve výpočtu zvýhodněna – prostřední graf na obr. 3
c) graf, ve kterém původní klíčová slova dostala smyčku, tyto smyčky a hrany vedoucí mezi klíčovými slovy dostaly váhu jedna a ostatní hrany dostaly váhu snižující se dle času jejich vzniku v grafu, tj. klíčová slova byla zvýhodněna a nově přidané termíny znevýhodněny – pravý graf na obr. 3.

Ovšem mohou být testovány i další varianty grafů.



Obr. 3: Typy použitých grafů (římská čísla zastupují dobu přidání vrcholu do grafu).

Při použití v Rocchio klasifikátoru (Manning et al., 2008) dosahovala nejlepších výsledků varianta grafu s označením *c*, pokud jsme klasifikátor trénovali na malé množině dokumentů (např. pouze 10 dokumentů pro 1 kategorii) - zde přesahovala i statistické metody trénované na stovkách dokumentů, viz (Dostal et al., 2014; Nykl et al., 2013). Pokud ale byla množina trénovacích dokumentů větší, tak docházelo k přetrénování a poklesu kvality klasifikace, což je problém, který bychom chtěli do budoucna vyřešit. Výhodou našeho přístupu tedy je, že dosahuje dobrých výsledků,

pokud je použita malá množina trénovacích dokumentů, a proto může být využit v situacích, kdy nemáme možnost získat větší množinu dokumentů pro trénování. Druhou výhodou je, že výsledky jsou v člověkem čitelné podobě a mohou být použity neprofesionálními uživateli.

V této oblasti jsme publikovali dva články (Dostal et al., 2014; Nykl et al., 2013) zabývající se využitím Linked Data a PageRanku pro klasifikaci a článek (Dostal et al., 2013), který popisuje, jak lze podobný princip využít pro pojmenování shluků.

Soupis vlastních publikací

Publikace zmíněné v následujícím soupisu vlastních publikací lze nalézt v příloze. Řazení publikací v příloze odpovídá řazení v tomto soupisu.

1) NYKL, Michal & JEŽEK, Karel. Varianty použití PageRanku pro citační analýzu. In: *DATAKON 2012*. Mikulov: Technická univerzita v Košiciach, 2012, pp. 87-97. ISBN: 978-80-553-1049-7.

2) HELLER, Petr, NYKL, Michal & JEŽEK, Karel. PageRank and analysis of citation cycles. In: *Proceedings of the Conference on Theory and Practice of Information Technologies*. Vrátna Dolina: CEUR-WS.org, 2011, pp. 89-90. ISBN: 978-80-89557-01-1.

3) NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin & FIALA, Dalibor. Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha: IADIS Press, 2013, pp. 61-64. ISBN: 978-972-8939-94-6.

4) DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. On-line: JATIT & LLS, vol. 53, no. 3, 2013, pp. 340-345. ISSN: 1992-8645.

5) DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel. Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII*. Praha: Springer International Publishing, 2014, pp. 37-43. ISBN: 978-3-319-01570-5.

4 Navrhovaný plán práce

Primárním cílem disertace bude použití PageRanku pro objektivní hodnocení autorů vědeckých publikací. První úlohou bude vyhodnocení záznamů z Web of Science (WoS). Zde nás zajímá, jak se změní výsledné pořadí autorů, pokud je PageRank aplikován buď na citační graf autorů, nebo na citační graf publikací, ze kterého jsou následně hodnoty publikací na autory přeneseny. Dále bychom chtěli zjistit, jaký vliv na výsledné pořadí autorů mají samocitace a váhy hran, které v citačním grafu autorů vyjadřují počet citací nebo berou v úvahu počet spoluautorů publikací. Použity budou záznamy WoS z let 1996 až 2005, které se nacházejí v Journal Citation Report (JCR) 2009 a jsou z oblasti počítačových věd. Pro porovnání výsledných pořadí autorů budou použita významná ocenění *ACM A.M. Turing Award* a *ACM SIGMOD E.F. Codd Innovations Award*, vysoce citované osoby ISI (*ISI Highly Cited*) a významní členové ACM (*ACM Fellows*). Přístup je podobný přístupu, který jsme použili v předchozím článku (Nykl & Ježek, 2012) pro vyhodnocení grafů vzniklých z dat databází DBLP (2004) a CiteSeer (2005).

Druhou úlohou bude navržení vyhodnocení autorů, které bude kombinovat Impact Factor časopisů a PageRank. Zde bychom chtěli otestovat několik variant vložení Impact Factoru do vzorce PageRanku (např. využití personalizace nebo vah hran), který bude aplikován na citační graf publikací a hodnoty publikací několika způsoby (rovnoměrně/nerovnoměrně) přeneseny na autory. Výsledky bychom navíc chtěli porovnat s dalšími čtyřmi vyhodnoceními a to s:

a) PageRankem, který místo Impact Factoru použije hodnoty PageRanku časopisů vypočítané z citačního grafu časopisů;
b) PageRankem, který nepoužívá personalizaci při hodnocení publikací;
c) PageRankem, který jako personalizaci použije h-index, aplikovaným na citační graf autorů;
d) PageRankem, který nepoužívá personalizaci, aplikovaným na citační graf autorů.

Varianty b) a d) budou sloužit jako referenční a varianty a) a c) by měly ukázat, zda lepší pořadí autorů nezískáme tak, že do vyhodnocení zahrneme h-index, či PageRank časopisů. Použita budou data z WoS a významná ocenění. Naše hypotézy jsou:

- Impact Factor doplněný do PageRanku hodnotícího publikace by měl zohlednit kvalitu časopisů, ve kterých byly publikace publikovány, a tím zlepšit výsledné pořadí autorů.
- nahrazení Impact Factoru PageRankem časopisů by mělo poskytnout nejlepší výsledné pořadí autorů, protože PageRank zohledňuje prestiž a Impact Factor „pouze“ popularitu.
- doplnění h-indexu do PageRanku hodnotícího citační graf autorů více zohlední vliv významnosti publikací při hodnocení autorů (očekáváme, že výsledky budou lepší, než když se tato personalizace nepoužije, ale horší, než výsledky získané využitím PageRanku časopisů).

Dalšími úlohami mohou být:

a) Porovnání WoS, CiteSeer a DBLP z pohledu citační analýzy – zde bychom chtěli porovnat výsledná pořadí autorů, získaná využitím PageRanku, s udílenými oceněními (zmíněnými např. v prvním odstavci této kapitoly) a určit, která z daných databází je lépe použitelná pro hodnocení autorů.
b) Znázornění vývoje významu autorů zjištěného využitím PageRanku a predikce vývoje budoucího – cílem je vytvořit diagramy pro několik nejvýznamnějších autorů zjištěných z WoS

využitím PageRanku, které budou demonstrovat jejich vývoj v průběhu několika let. Použito bude pravděpodobně tříleté časové okénko, dle kterého z dat WoS vznikne osm kolekcí (1996-1998, 1997-1999, … , 2003-2005), ve kterých budou PageRankem nalezeni významní autoři. Pro autory, kteří se budou vyskytovat mezi prvními sedmi autory v libovolné kolekci, budou vytvořeny diagramy znázorňující vývoj jejich významnosti v daných letech. Cílem bude predikovat budoucí významnost autorů na základě jejich předchozího vývoje. Pozn.: která varianta PageRanku bude použita, zda bude penalizován čas vzniku publikace a jak velké časové okénko bude použito, jsou aktuálně otevřené otázky.

Sekundárním cílem bude výzkum aplikovatelnosti PageRanku v dalších oblastech. Stěžejní oblastí bude volba vlastností využitím ontologie zastoupené Linked Data. K dispozici máme data Linked Data z DBpedia.org, kolekci *20 News groups* (manuálně klasifikované novinové články) a vlastní kolekci konferenčních *Call for Paper* (CFP). Každý dokument je navíc předzpracován TF-IDF či χ^2 a nalezené vlastnosti/termíny namapovány do Linked Data. Naším úkolem je nalezení takové vlastnosti (či vlastností), která nejvíce vystihuje téma dokumentu, k čemuž používáme PageRank. Nalezené základní vlastnosti navíc můžeme využitím Linked Data rozšířit o obecnější vlastnosti, které dokument původně neobsahoval. To nám umožňuje dokumenty, které obsahovaly např. termíny *MySQL* a *IBM DB2* zařadit do kategorie *Relační databázové systémy*. Tento postup může být využit v klasifikaci, shlukování nebo štítkování. V následující práci bychom rádi vyzkoušeli další varianty grafů vlastností, které by mohly dosahovat lepších výsledků, než varianty, které jsme již testovali v (Dostal et al., 2014; Nykl et al., 2013).

Další oblastí, ve které chceme aplikovat PageRank, je analýza provázanosti firem na základě zainteresovaných osob (cílem je určení firem či osob, které nejvíce ovlivňují ostatní firmy). Také bychom chtěli využitím PageRanku predikovat vítěze sportovních utkání (zde očekáváme, že PageRank poskytne lepší predikci, než tabulka s aktuálním postavením týmů v soutěži nebo tabulka z předešlé sezóny). Úkoly v těchto oblastech ale zatím přenecháme spíše na své bakaláře či diplomanty.

Harmonogram práce

Následující harmonogram znázorňuje předpokládaný plán práce, který se ale v průběhu vypracovávání může změnit z důvodu nalezení lukrativnější oblasti, ve které by PageRank mohl být aplikován, či z důvodu detailnějšího rozpracování některé ze zmíněných úloh:

Do ledna 2014	• Vyhodnocení autorů ve WoS a publikování výsledků (obdoba článku (Nykl & Ježek, 2012)). • Publikování současného stavu hodnocení významnosti časopisů, autorů a publikací.
Do května 2014	• Zakomponování Impact Factoru, PageRanku časopisů a h-indexu do PageRanku, vyhodnocení autorů ve WoS a publikování výsledků.
Do října 2014	• Vytvoření lepší varianty volby vlastností a publikování výsledků.
Do března 2015	• Znázornění a predikce vývoje významnosti autorů ve WoS nebo porovnání WoS, DBLP a CiteSeeru a publikování výsledků.
Do července 2015	• Text disertační práce

Literatura

AMARA, Nabil & LANDRY, Réjean: Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science. *Scientometrics*. 2012, vol. 93, no. 3, pp. 553–581. ISSN 0138-9130.

ASSIMAKIS, N. & ADAM, M.: A new author's productivity index: p-index. *Scientometrics*. 2010, vol. 85, no. 2, pp. 415–427. ISSN 01389130.

AVRACHENKOV, Konstantin, DOBRYNIN, Vladimir, NEMIROVSKY, Danil, PHAM, Son Kim & SMIRNOVA, Elena: Pagerank based clustering of hypertext document collections. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*. New York, USA: ACM Press, 2008, p. 873–874. ISBN 9781605581644.

BARABÁSI, Albert-László: *V pavučině sítí*. Praha: Paseka, 2005. ISBN 8071857513.

BAR-ILAN, Judit: Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*. 2007, vol. 74, no. 2, pp. 257–271. ISSN 0138-9130.

BELLIS, Nicola De: *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, Toronto, Plymouth: The Scarecrow Press, 2009. ISBN 9780810867130.

BENZI, Michele & KUHLEMANN, Verena: *Chebyshev acceleration of the GeneRank algorithm*. 2012.

BERCHENKO, Yakir, DALIOT, Or & BRUELLER, Nir N.: Intra-firm information flow: a content-structure perspective. In: *Advances in Intelligent Data Analysis X*. Porto: Springer Berlin Heidelberg, 2011, p. 34–42.

BOLDI, Paolo, BONCHI, Francesco, CASTILLO, Carlos & VIGNA, Sebastiano: Voting in social networks. In: *CIKM'09*. New York, USA: ACM Press, 2009. ISBN 9781605585123.

BOLLEN, Johan, RODRIQUEZ, Marko A. & VAN DE SOMPEL, Herbert: Journal status. *Scientometrics*. 2006, vol. 69, no. 3, pp. 669–687.

BONACICH, Phillip: Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*. 1972, vol. 2, no. 1, pp. 113–120. ISSN 0022-250X.

BORODIN, Allan, ROBERTS, Gareth O., ROSENTHAL, Jeffrey S. & TSAPARAS, Panayiotis: Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*. 2005, vol. 5, no. 1, pp. 231–297. ISSN 15335399.

BRIN, Sergey & PAGE, Lawrence: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998, vol. 30, no. 1-7, pp. 107–117. ISSN 01697552.

DI CARO, Luigi, CATALDI, Mario & SCHIFANELLA, Claudio: The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*. 2012, vol. 93, no. 3, pp. 583–607. ISSN 0138-9130.

DING, Chris, HE, Xiaofeng, HUSBANDS, Parry, ZHA, Hongyuan & SIMON, Horst: PageRank, HITS and a unified framework for link analysis. In: *Proceedings of the 25th annual international ACM SIGIR*

conference on Research and development in information retrieval - SIGIR '02. New York, USA: ACM Press, 2002, p. 249–253. ISBN 1581135610.

DING, Ying & YAN, Erjia: PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*. 2009, vol. 60, no. 11, pp. 2229–2243.

DING, Ying: Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*. 2011, vol. 62, no. 2, pp. 236–245.

DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel: Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. 2013, vol. 53, no. 3, pp. 340–345.

DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel: Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII*. Praha: Springer International Publishing, 2014, p. 37–43.

EGGHE, Leo: The Hirsch-Index and Related Impact Measures. *Annual Review of Information Science and Technology*. 2010, vol. 44, no. 1, pp. 65–114.

ELKINS, Mark R., MAHER, Christopher G., HERBERT, Robert D., MOSELEY, Anne M. & SHERRINGTON, Catherine: Correlation between the Journal Impact Factor and three other journal citation indices. *Scientometrics*. 2010, vol. 85, no. 1, pp. 81–93. ISSN 01389130.

ERKAN, Günes & RADEV, Dragomir R.: LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*. 2004, vol. 22, pp. 457–479.

FERRARA, Emilio: *Mining and Analysis of Online Social Networks*. Messina, 2012. University of Messina.

FIALA, Dalibor, ROUSSELOT, François & JEŽEK, Karel: PageRank for bibliographic networks. *Scientometrics*. 2008, vol. 76, no. 1, pp. 135–158. ISSN 0138-9130.

FIALA, Dalibor: Bibliometric analysis of CiteSeer data for countries. *Information Processing & Management*. 2012, vol. 48, no. 2, pp. 242–253. ISSN 03064573.

FIALA, Dalibor: Mining citation information from CiteSeer data. *Scientometrics*. 2011, vol. 86, no. 3, pp. 1–12.

FIALA, Dalibor: Time-aware PageRank for bibliographic networks. *Journal of Informetrics*. 2012, vol. 6, no. 3, pp. 370–388. ISSN 17511577.

FRANCESCHINI, Fiorenzo, MAISANO, Domenico & MASTROGIACOMO, Luca: The effect of database dirty data on h-index calculation. *Scientometrics*. 2013, vol. 95, no. 3, pp. 1179–1188. ISSN 0138-9130.

FREEMAN, Linton C.: A set of measures of centrality based on betweenness. *Sociometry*. 1977, vol. 40, no. 1, pp. 35–41.

FREEMAN, Linton C.: Centrality in social networks conceptual clarification. *Social networks*. 1979, vol. 1, pp. 215–239.

GARFIELD, Eugene: Citation analysis as a tool in journal evaluation. *Science*. 1972, vol. 178, no. 60, pp. 471–479.

GARFIELD, Eugene: Citation indexes for science: A new dimension in documentation through association of ideas. *Science*. 1955, vol. 122, pp. 108–111.

GILES, C. Lee, BOLLACKER, Kurt D. & LAWRENCE, Steve: CiteSeer: an automatic citation indexing system. In: *Proceedings of the third ACM conference on Digital libraries - DL '98*. New York: ACM Press, 1998, p. 89–98. ISBN 0897919653.

HADDOW, Gaby & GENONI, Paul: Citation analysis and peer ranking of Australian social science journals. *Scientometrics*. 2010, vol. 85, no. 2, pp. 471–487. ISSN 01389130.

HAN, Yo-Sub, KIM, Laehyun & CHA, Jeong-Won: Computing user reputation in a social network of web 2.0. *Computing and Informatics*. 2012, vol. 31, pp. 1001–1016.

HANNEMAN, Robert A. & RIDDLE, Mark: *Introduction to social network methods*. 2005 [accessed. 18. December 2012]. Retrieved from: http://faculty.ucr.edu/~hanneman/nettext/

HAO, Fei, PEI, Zheng, ZHU, Chunsheng, WANG, Guojun & YANG, Laurence T.: User attractor: An operator for the evaluation of social influence. *Future Generation Computer Systems*. 2012. ISSN 0167739X.

HARZING, Anne-Wil: A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*. 2013, vol. 94, no. 3, pp. 1057–1075. ISSN 0138-9130.

HELLER, Petr, NYKL, Michal & JEŽEK, Karel: PageRank and analysis of citation cycles. In: *ITAT 2011*. Vrátna Dolina (SVK): CEUR-WS.org, 2011, p. 89–90.

HIRSCH, J. E.: An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*. 2005, vol. 102, no. 46, pp. 16569–16572.

HO, Yuh-Shan: The top-cited research works in the Science Citation Index Expanded. *Scientometrics*. 2013, vol. 94, no. 3, pp. 1297–1312. ISSN 0138-9130.

CHIRITA, Paul Alexandru, NEJDL, Wolfgang, SCHLOSSER, Mario & SCURTU, Oana: Personalized Reputation Management in P2P Networks. In: *ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*. Hiroshima, JP: CEUR-WS.org, 2004.

JACSÓ, Péter: The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*. 2011, vol. 35, no. 6, pp. 983–997. ISSN 1468-4527.

KENDALL, M. G.: A new measure of rank correlation. *Biometrika*. 1938, vol. 30, no. 1, pp. 81–93.

KLEINBERG, Jon M.: Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 1999, vol. 46, no. 5, pp. 604–632. ISSN 00045411.

KURTZ, Michael J. & BOLLEN, Johan: Usage Bibliometrics. *Annual Review of Information Science and Technology*. 2011, vol. 44, no. 1, pp. 3–64.

LANGVILLE, Amy N. & MEYER, Carl D.: *Google's PageRank and Beyond The Science of Search Engine Rankings*. Princeton, NJ, USA: Princeton University Press, 2006. ISBN 0691122024.

LEE, Tim-Berners: *Linked Data - Design Issues*. 2006 [accessed. 20. September 2013]. Retrieved from: http://www.w3.org/DesignIssues/LinkedData.html

LEMPEL, Ronny & MORAN, Shlomo P.: SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*. 2001, vol. 19, no. 2, pp. 131–160. ISSN 10468188.

LEMPEL, Ronny & MORAN, Shlomo P.: The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*. 2000, vol. 33, no. 1-6, pp. 387–401. ISSN 13891286.

LEYDESDORFF, Loet: An evaluation of impacts in "Nanoscience & nanotechnology": steps towards standards for citation analysis. *Scientometrics*. 2013, vol. 94, no. 1, pp. 35–55. ISSN 0138-9130.

LIBEN-NOWELL, David & KLEINBERG, Jon: The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*. 2007, vol. 58, no. 7, pp. 1019–1031. ISSN 15322882.

LIU, Xiaoming, BOLLEN, Johan, NELSON, Michael L. & VAN DE SOMPEL, Herbert: Co-Authorship Networks in the Digital Library Research Community. *Information Processing & Management*. 2005, vol. 41, no. 6, p. 28.

LÓPEZ, Roque Enfique, BARREDA, Dennis, TEJADA, Javier & CUADROS, Ernesto: MFSRank: an unsupervised method to extract keyphrases using semantic information. In: *Advances in Artificial Intelligence*. Puebla: Springer Berlin Heidelberg, 2011, p. 338–344.

MA, Nan, GUAN, Jiancheng & ZHAO, Yi: Bringing PageRank to the citation analysis. *Information Processing & Management*. 2008, vol. 44, no. 2, pp. 800–810. ISSN 03064573.

MANNING, Christopher D., RAGHAVAN, Prabhakar & SCHÜTZE, Hinrich: *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 0521865719.

MIHALCEA, Rada & TARAU, Paul: TextRank: Bringing order into texts. *Proceedings of EMNLP*. 2004.

MINGERS, John & LIPITAKIS, Evangelia: Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*. 2010, vol. 85, no. 2, pp. 613–625. ISSN 01389130.

MOED, Henk F.: Citation Analysis in Research Evaluation. In: *Information Knowledge and Science Management*. Dordrecht, NL: Springer, 2005, p. 333. ISBN 1402037139.

MORRISON, Julie L, BREITLING, Rainer, HIGHAM, Desmond J & GILBERT, David R: GeneRank: using search engine technology for the analysis of microarray experiments. *BMC bioinformatics*. 2005, vol. 6, no. 1, pp. 233–247. ISSN 1471-2105.

MRYGLOD, O., KENNA, R., HOLOVATCH, Yu. & BERCHE, B.: Absolute and specific measures of research group excellence. *Scientometrics*. 2013, vol. 95, no. 1, pp. 115–127. ISSN 0138-9130.

NEWMAN, M. E. J.: *Networks: An Introduction*. New York, USA: Oxford University Press, 2010. ISBN 9780470749838.

NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin & FIALA, Dalibor: Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha: IADIS Press, 2013.

NYKL, Michal & JEŽEK, Karel: Varianty použití PageRanku pro citační analýzu. In: *DATAKON 2012*. Mikulov, CZE: Technická univerzita v Košiciach, 2012, p. 87–97.

NYKL, Michal: *Vyhodnocování informačních sítí*. Plzeň, 2011. Západočeská univerzita v Plzni.

ORTIZ, Roberto & PINTO, David: BUAP: An unsupervised approach to automatic keyphrase extraction from scientific articles. In: *Proceedings of the 5th international workshop on semantic evaluation*. Uppsala, SWE: Association for Computational Linguistics, 2010, p. 174–177.

PAGE, Lawrence, BRIN, Sergey, MOTWANI, Rajeev & WINOGRAD, Terry: *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford: Stanford InfoLab. 1999

RUHNAU, Britta: Eigenvector-centrality — a node-centrality? *Social Networks*. 2000, vol. 22, no. 4, pp. 357–365. ISSN 03788733.

RYJÁČEK, Zdeněk: *Teorie grafů a diskrétní optimalizace 1*. Plzeň, 2001. Západočeská univerzita v Plzni.

SAYYADI, Hassan & GETOOR, Lise: FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In: *The Ninth SIAM International Conference on Data Mining*. Nevada: SIAM, 2009, p. 533–544.

SIDIROPOULOS, Antonis & MANOLOPOULOS, Yannis: A citation-based system to assist prize awarding. *ACM SIGMOD Record*. 2005, vol. 34, no. 4, pp. 54–60. ISSN 01635808.

SIDIROPOULOS, Antonis & MANOLOPOULOS, Yannis: A new perspective to automatically rank scientific conferences using digital libraries. *Information Processing & Management*. 2005, vol. 41, no. 2, pp. 289–312. ISSN 03064573.

SIDIROPOULOS, Antonis & MANOLOPOULOS, Yannis: Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*. 2006, vol. 79, no. 12, pp. 1679–1700. ISSN 01641212.

SILER, Kyle: Citation choice and innovation in science studies. *Scientometrics*. 2012, vol. 95, no. 1, pp. 385–415. ISSN 0138-9130.

SMIRNOVA, Elena, AVRACHENKOV, Konstantin & TROUSSE, Brigitte: Using Web Graph Structure for Person Name Disambiguation. *Third Web People Search Evaluation Forum (WePS-3), CLEF*. 2010.

SPEARMAN, C.: The proof and measurement of association between two things. *The American journal of psychology*. 1904, vol. 15, no. 1, pp. 72–101.

ÚŘAD VLÁDY ČR: *Metodika hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů (platná pro léta 2010 a 2011 a rok 2012)*. 2012

ÚŘAD VLÁDY ČR: *Metodika hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů (platná pro léta 2013 až 2015)*. 2013

VOEVODSKI, Konstantin, TENG, Shang-Hua & XIA, Yu: Spectral affinity in protein networks. *BMC systems biology*. 2009, vol. 3, no. 1, pp. 112–125. ISSN 1752-0509.

WALKER, Dylan, XIE, Huafeng, YAN, Koon-Kiu & MASLOV, Sergei: Ranking Scientific Publications Using a Simple Model of Network Traffic. *Journal of Statistical Mechanics: Theory and Experiment*. 2006, vol. 2007, no. 06, p. 4.

WANG, G. Alan, JIAO, Jian, ABRAHAMS, Alan S., FAN, Weiguo & ZHANG, Zhongju: ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*. 2013, vol. 54, no. 3, pp. 1442–1451. ISSN 01679236.

WANG, Haijun, LIU, Minyan, HONG, Song & ZHUANG, Yanhua: A historical review and bibliometric analysis of GPS research from 1991–2010. *Scientometrics*. 2012, vol. 95, no. 1, pp. 35–44. ISSN 0138-9130.

WEST, Jevin D.: *Eigenfactor: ranking and mapping scientific knowledge*. Washington, USA, 2010. University of Washington.

XING, Wenpu & GHORBANI, Ali: Weighted PageRank algorithm. In: *Proceedings of the Second Annual Conference on Communication Networks and Services Research*. Fredericton, CA: IEEE, 2004, p. 305–314. ISBN 0-7695-2096-0.

YAN, Erjia, DING, Ying & SUGIMOTO, Cassidy R.: P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*. 2010, vol. 62, no. 3, pp. 467–477.

YAN, Erjia & DING, Ying: Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*. 2009, vol. 60, no. 10, pp. 2107–2118. ISSN 15322882.

YAN, Erjia & DING, Ying: Discovering author impact: A PageRank perspective. *Information Processing & Management*. 2011, vol. 47, no. 1, pp. 125–134. ISSN 03064573.

YAN, Erjia & DING, Ying: Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*. 2012, vol. 63, no. 7, pp. 1313–1326.

YAN, Erjia & DING, Ying: Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*. 2010, vol. 61, no. 8, pp. 1635–1643.

YANG, Zaihan, HONG, Liangjie & DAVISON, Brian D.: Topic-driven multi-type citation network analysis. In: *RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information*. 2010, p. 24–31.

YU, Kun, CHEN, Xiaobing & CHEN, Jianhong: A multidimensional PageRank algorithm of Literatures. *Journal of Theoretical and Applied Information Technology*. 2012, vol. 44, no. 2, pp. 308–315.

ZHAO, Dangzhi: Going beyond counting first authors in author co-citation analysis. *Proceedings of the American Society for Information Science and Technology*. 2005, vol. 42, no. 1.

ZHOU, Ding, ORSHANSKIY, Sergey A., ZHA, Hongyuan & GILES, C. Lee: Co-ranking authors and documents in a heterogeneous network. In: *Seventh IEEE International Conference on Data Mining*. Omaha, USA: IEEE, 2007, p. 739–744.

ZHU, Wenjia & GUAN, Jiancheng: A bibliometric study of service innovation research: based on complex network analysis. *Scientometrics*. 2013, vol. 94, no. 3, pp. 1195–1216. ISSN 0138-9130.

Příloha

Příloha obsahuje texty těchto článků:

1) NYKL, Michal & JEŽEK, Karel. Varianty použití PageRanku pro citační analýzu. In: *DATAKON 2012*. Mikulov: Technická univerzita v Košiciach, 2012, pp. 87-97. ISBN: 978-80-553-1049-7.

2) HELLER, Petr, NYKL, Michal & JEŽEK, Karel. PageRank and analysis of citation cycles. In: *Proceedings of the Conference on Theory and Practice of Information Technologies*. Vrátna Dolina: CEUR-WS.org, 2011, pp. 89-90. ISBN: 978-80-89557-01-1.

3) NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin & FIALA, Dalibor. Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha: IADIS Press, 2013, pp. 61-64. ISBN: 978-972-8939-94-6.

4) DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. On-line: JATIT & LLS, vol. 53, no. 3, 2013, pp. 340-345. ISSN: 1992-8645.

5) DOSTAL, Martin, NYKL, Michal & JEŽEK, Karel. Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII*. Praha: Springer International Publishing, 2014, pp. 37-43. ISBN: 978-3-319-01570-5.