Varianty použití PageRanku pro citační analýzu

Michal NYKL, Karel JEŽEK

Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni
Univerzitní 22, 306 14 Plzeň
nyklm@kiv.zcu.cz, jezek_ka@kiv.zcu.cz

Abstrakt. V článku si představíme problematiku vyhodnocování citačních sítí, jejímž cílem je určení významnosti autorů. Ukážeme si vyhodnocování citačních sítí algoritmem PageRank s ohledem na spoluautorství a vypočtená pořadí autorů porovnáme s nejznámějšími oceněními jako např. ACM A. M. Turing Award, ACM Fellows a další. Analýzu citační sítě lze využít např. při přidělování grantů, či při hledání osob na vedoucí pozice ve výzkumných institucích.

Klíčová slova: PageRank, citační analýza, spoluautorství, CiteSeer, DBLP, Turing Award, Codd Award, ACM Fellows, ISI Highly Cited.

1 Úvod

V roce 1998 Page a Brin představili algoritmus PageRank sloužící k určení významnosti webových stránek na základě hypertextových odkazů a významnosti webových stránek, ze kterých dané hypertextové odkazy vedou, viz [8]. Algoritmus je aplikován na síť, kde uzly jsou webové stránky a hrany vyjadřují, že z jedné stránky vede odkaz na stránku jinou (váha všech hran v síti je 1). Algoritmus iterativním způsobem ohodnotí jednotlivé uzly sítě a tím umožní vytvořit z ohodnocených uzlů žebříček, tedy seřadit uzly od nejvýznamnějšího k nejméně významnému.

Od svého vzniku byl algoritmus PageRank dále analyzován, upravován [6] a aplikován na rozličné sítě (mimo jiné na síť autorských citací, viz dále) za účelem určení významnosti jednotlivých uzlů těchto sítí.

Naší snahou je zjistit, jaký vliv má uvažování spoluautorství v citační analýze využívající algoritmus PageRank. Za tímto účelem ukážeme několik variant sítí autorských citací, ve kterých jsme různými způsoby zohlednili spoluautorství. Vyhodnotíme je pomocí algoritmu PageRank a získané žebříčky následně porovnáme s některými vědeckými institucemi udělovanými oceněními. Naším cílem je nalézt takové varianty vyhodnocení citačních sítí, které oceněným vědcům přiřadí největší PageRankové skóre.

K vyhodnocením různých variant úprav sítě autorských citací navíc přidáme metody, ve kterých určíme významnost publikací z jejich citační sítě a tuto významnost poté přeneseme na autory.

2 Sociální sítě autorů

Jednou z prvních vyhodnocovaných sociálních sítí autorů, která vznikla z jejich publikační činnosti, je síť autorských citací. V této síti jsou uzly, představující jednotlivé autory, provázány hranou vždy, když autor ve své publikaci citoval publikaci jiného autora (pozn.:

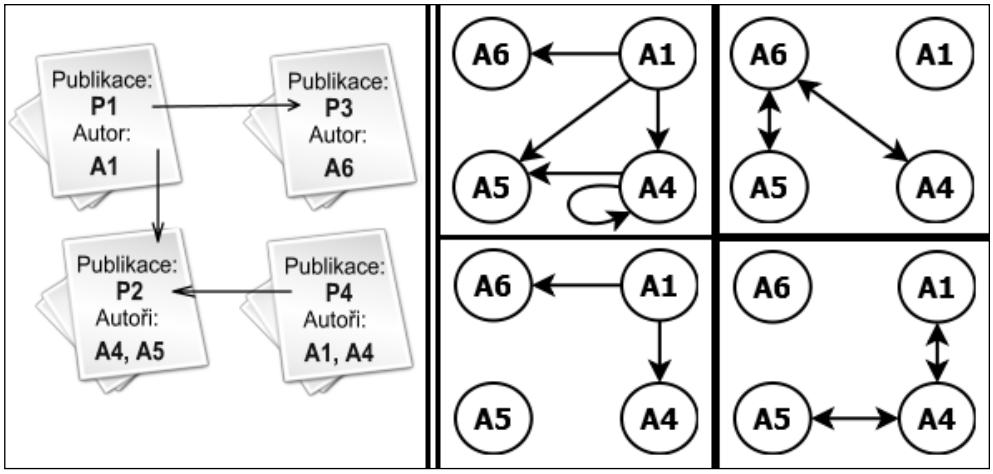
v síti tedy vznikla hrana od prvního autora k druhému). Různé varianty vyhodnocení nalezneme např. v [1], [3], [5] a [7].

Další síť, která též využívá citací, je síť společně-citovaných autorů. Zde jsou uzly sítě provázány hranou, pokud autoři byli citováni jiným autorem v jedné jeho publikaci, tj. pokud autor A citoval publikaci autora B a publikaci autora C, tak mezi autory B a C povede hrana. Vyhodnocení této sítě můžeme vidět např. v [11].

Důležitým aspektem při vytváření sítí autorů dle citací je, zda vytváříme síť ze všech autorů publikace, tzv. All-authors, nebo pouze z autorů, kteří jsou v publikaci uvedeni na prvním místě, tzv. First-authors, viz [11]. Dále v tomto článku se zaměříme pouze na variantu All-authors.

Jinou alternativou sociální sítě autorů je síť spoluautorů, kde mezi autory vede hrana, pokud společně napsali alespoň jednu publikaci. Související vyhodnocení např. viz [7] a [10].

Příklady výše zmíněných sítí můžeme vidět na obr. 1. Sítě autorů v pravé části obrázku byly vytvořeny z citační sítě publikací, která je v části levé.



Obr. 1. Několik alternativ sítí
(vlevo citační síť publikací, uprostřed sítě autorských citací – nahoře All-authors a dole First-authors, vpravo nahoře síť společně-citovaných autorů All-authors, vpravo dole síť spoluautorů).

3 Vyhodnocení citačních sítí s ohledem na spoluautorství

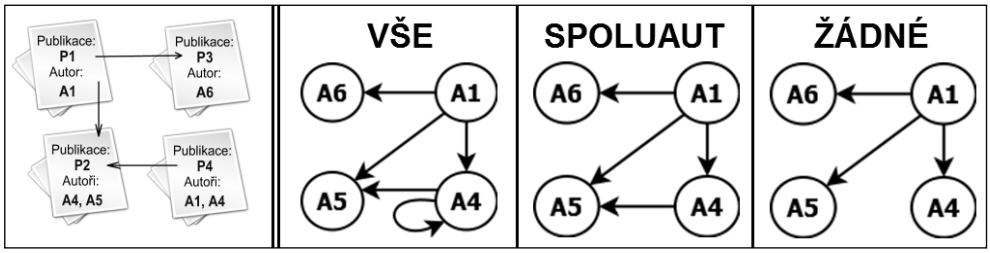
Naším cílem v této části článku bude ukázat si několik možností vytvoření citační sítě autorů s ohledem na spoluautorství. Nejprve se podíváme na to, jak lze zohlednit samocitace (tedy fakt, že autor cituje své publikace) při tvorbě sítě, poté si ukážeme, jakým způsobem můžeme přiřadit váhy hranám v síti autorských citací a jaký algoritmus, potažmo vzorec PageRanku použít pro vyhodnocení sítě, abychom získali ohodnocení autorů. Dále si představíme dvě varianty vyhodnocení, které pracují s ohodnocenými publikacemi (získáno vyhodnocením citační sítě publikací) a převádí jejich PageRanková skóre na autory. V závěru této části se podíváme na některé další alternativy vyhodnocení sítě.

3.1 Tvorba citačních sítí s různými variantami samocitací

Jedním z aspektů při vytváření sítě autorských citací je, jak moc jsme ochotni připustit, aby významnost autora záležela i na jeho citování sama sebe, tj. na jeho samocitacích. V tomto případě se lze zachovat třemi způsoby:

a) samocitace uznáme jako plnohodnotné citace (značeno **VŠE**),
b) odstraníme citace mezi publikacemi se stejným autorem (značeno **ŽÁDNÉ**),
c) pokud autor cituje svou publikaci, pak cituje pouze své spoluautory v této publikaci, ale necituje sebe (značeno **SPOLUAUT**).

Příklady těchto sítí viz obr. 2. (pozn.: v případě VŠE můžeme posléze penalizovat váhu hran představujících samocitace, např. viz [9]).



Obr. 2. Různé varianty zohlednění samocitací v síti autorských citací.

3.2 Určení vah hranám v síti autorských citací

Váhu hran v síti autorských citací lze, s ohledem na spoluautorství, určit mnoha způsoby, např. viz [3]. My zde nyní porovnáme způsoby tři a to:

i) každá hrana v citační síti autorů má váhu jedna (značeno **1**),
ii) váha hrany vyjadřuje, kolikrát autor citoval publikace daného autora (značeno **N**),
iii) pokud autor citoval publikaci s N autory, tak z jeho uzlu povede na uzel každého z těchto autorů hrana s váhou 1/N (značeno **1/N**). Souhlasné hrany mezi dvěma autory se spojí a jejich váhy se sečtou.

Přiřazení vah hranám v sítích autorských citací, které jsou ukázány na obr. 2, viz tab. 1. Měli bychom upozornit na fakt, že pokud z uzlu vede pouze jedna výstupní hrana, tak při vyhodnocení algoritmem PageRank nezáleží na váze této hrany, protože uzel po dané hraně předá celé své PageRankové skóre, viz hrana **(A4, A5)** ve sloupci **1/N** sítě **SPOLUAUT**.

hrany	VŠE			SPOLUAUT			ŽÁDNÉ		
	1	**N**	**1/N**	**1**	**N**	**1/N**	**1**	**N**	**1/N**
(A1, A6)	1	1	1	1	1	1	1	1	1
(A1, A5)	1	2	1	1	2	1	1	1	0,5
(A1, A4)	1	2	1	1	2	1	1	1	0,5
(A4, A5)	1	1	0,5	1	1	0,5	---	---	---
(A4, A4)	1	1	0,5	---	---	---	---	---	---

Tab. 1. Možnosti přiřazení vah hranám v síti autorských citací.

3.3 Vyhodnocení sítě algoritmem PageRank

Pro vyhodnocování vzniklých citačních sítí s rozličným ohodnocením hran bude použit vzorec PageRanku, viz (1), kde $P_x(A)$ je skóre PageRanku uzlu A v iteraci x, d je damping faktor (obvykle 0,85, viz [6]), $|V|$ je počet uzlů v síti, U je množina uzlů, z nichž vede hrana na uzel A, w_{utoA} je váha hrany z uzlu u do uzlu A, w_{uout} je součet vah výstupních hran z uzlu u a D je množina uzlů bez výstupní hrany. Vzorec ošetřuje uzly bez výstupní hrany (tzv. dangling nodes) tak, že jim pomyslně doplní hrany s váhou 1 na všechny uzly v síti i na sebe sama, viz poslední zlomek vzorce (1). Vzorec (1) budeme v tabulkách značit jako **Neupr.**

Jak z popisu vyplývá, vzorec PageRanku je iterační a výpočet tedy končí po zvoleném počtu iterací, či pokud se skóre PageRanku jednotlivých uzlů mezi dvěma iteracemi neliší více, než zvolené ε, nebo pokud se mezi iteracemi již nemění pořadí uzlů tvořené dle jejich skóre PageRanku. My jsme zvolili ukončení po 50 iteracích.

$$P_{x+1}(A) = \frac{(1-d)}{|V|} + d * ((\sum_{u \in U} \frac{P_x(u) * w_{utoA}}{w_{uout}}) + \frac{\sum_{s \in D} P_x(s)}{|V|}) \quad (1)$$

První alternativou vzorce PageRanku, kterou při vyhodnocení také použijeme, je vzorec (2), v tabulkách značen **Upr.1**. V tomto vzorci je upraven první zlomek, představující zcela náhodný přístup na uzel, tak, že pravděpodobnost náhodného přístupu na uzel závisí na počtu publikací daného autora. To odpovídá skutečnosti, že čím více publikací autor napsal, tím ho lze ze statistického hlediska považovat za významějšího. A_{Pub} je počet publikací autora A, V je množina všech uzlů, zde autorů, a t_{Pub} je počet publikací autora t.

$$P_{x+1}(A) = \frac{(1-d) * \mathrm{A_{Pub}}}{\sum_{t \in V} \mathrm{t_{Pub}}} + d * ((\sum_{u \in U} \frac{P_x(u) * w_{utoA}}{|w_{uout}|}) + \frac{\sum_{s \in D} P_x(s)}{|V|}) \quad (2)$$

3.4 Ohodnocení autorů na základě hodnocení jejich publikací

Předchozí tři části ukazují, jak lze získat skóre PageRanku autorů ze sítě autorských citací. V této části se zaměříme na variantu, kdy pomocí PageRanku vyhodnotíme citační síť publikací, tj. použijeme vzorec (1), který na místo s autory bude pracovat s publikacemi (váhy všech hran v síti jsou rovny jedné), a poté předáme PageRanková skóre publikací jejich autorům. To provedeme dvěma způsoby. Buďto publikace rovnoměrně rozdělí svá skóre svým autorům, tedy dle vzorce (3), nebo každý autor získá součet skóre všech svých publikací, viz vzorec (4), a popř. se následně provede normalizace (což na žebříček autorů nemá vliv). Ve vzorcích A_{PUB} představuje množinu publikací, jejichž jedním z autorů je autor A, $P(b)$ je skóre PageRanku publikace b a b_{Aut} je počet autorů publikace b.

Vzorec (3) bude v tabulkách značen „**části**“ a vzorec (4) bude značen „**celé**“.

$$\mathrm{P(A)} = \sum_{b \in \mathrm{A_{PUB}}} \frac{\mathrm{P}(b)}{b_{\mathrm{Aut}}} \quad (3)$$

$$\mathrm{P(A)} = \sum_{b \in \mathrm{A_{PUB}}} \mathrm{P}(b) \quad (4)$$

Stejně, jako jsme při vyhodnocování sítě autorských citací použili alternativu Upr.1, která určuje náhodný přístup na uzel autora dle počtu jeho publikací, tak zde pro vyhodnocení citační sítě publikací použijeme alternativu vzorce PageRanku, která analogicky určuje náhodný přístup na uzel publikace dle počtu jejích autorů, viz vzorec (5).

Ve vzorci B_{Aut} představují počet autorů publikace *B*, *V* je opět množina všech uzlů, zde publikací, a b_{Aut} je počet autorů publikace *b*. Tuto alternativu vzorce PageRanku budeme v tabulkách značit **Upr.2**.

$$P_{x+1}(B) = \frac{(1-d)*\mathrm{B_{Aut}}}{\sum_{b\in V} \mathrm{b_{Aut}}} + d * ((\sum_{u\in U} \frac{P_x(u)*w_{u\mathrm{to}B}}{|w_{u\mathrm{out}}|}) + \frac{\sum_{s\in D} P_x(s)}{|V|}) \qquad (5)$$

3.5 Jiné možnosti vyhodnocení sítě

Algoritmus PageRank není jedinou možností, jak vyhodnotit síť (např. autorských citací). Další možností jsou např. tzv. míry centrality (*Centrality Measure*). K nim patří vážený/nevážený vstupní/výstupní stupeň uzlu (*Degree*), tj. počet hran uzlu, či součet vah těchto hran, blízkost (*Closeness*), tj. inverzní hodnota součtu nejkratších cest mezi zvoleným uzlem a všemi ostatními uzly v síti (lze říci, že blízkost vyjadřuje, jak dlouho se budou informace šířit ze zvoleného uzlu do všech ostatních uzlů sítě), a mezilehlost (*Betweenness*), která říká, na kolika nejkratších cestách mezi uzly sítě zvolený uzel leží (lze si přestavit jako úroveň řízení komunikace mezi uzly sítě). K mírám centrality lze přiřadit i algoritmus PageRank, který je zde chápán jako varianta vlastního vektoru (*Eigenvector*) představujícího míru vlivu jednotlivých uzlů v síti. Stručný popis měr centrality a jejich použití např. viz [10].

Jinou alternativou vyhodnocení sítě je algoritmus HITS (*Hypertext Induced Topic Search*), viz [4], který způsobem velmi podobným PageRanku přiřazuje každému uzlu sítě dvě skóre. První z nich vyjadřuje, jak je uzel dobrou autoritou (*authority*), tj. z kolika a jak hodnocených uzlů na zvolený uzel vede hrana. Druhé skóre zase říká, jak je uzel dobrým rozcestníkem (*hub*), tj. na kolik a jak hodnocených uzlů ze zvoleného uzlu vede hrana.

4 Použitá data

Citační síť se v praxi obvykle vytváří z bibliografické databáze, která obsahuje záznamy o publikacích, ke kterým jsou připojeny citace (tj. odkazy na v publikaci citované jiné publikace).

My pro náš experiment použijeme bibliografické databáze CiteSeer (2005) a DBLP (2004), které jsme zvolili na základě prací [2] a [5] a které se soustředí na oblast výzkumu v informačních vědách. Databáze CiteSeer je udržována strojově a ve verzi z roku 2005 obsahuje 648897 publikací, ve kterých bylo rozpoznáno 406465 různých autorů a 1542165 vzájemných citací mezi publikacemi. Databáze DBLP je udržována manuálně a ve verzi z roku 2004 obsahuje 469804 publikací, 315485 autorů a 100621 vzájemných citací. Za povšimnutí stojí, že v databázi CiteSeer připadá na každou publikaci průměrně 2,38 citací, kdežto v DBLP je to pouze 0,22 citací.

Ze zvolených bibliografických databází vytvoříme všechny sítě popsané v části 3 a vyhodnotíme je tamtéž zmíněnými postupy vyhodnocení. Získané žebříčky autorů poté porovnáme s udílenými oceněními ACM SIGMOD E. F. Codd Innovation Award *(značeno **CODD**, oceněných osob 19, v CiteSeer nalezeno 18, v DBLP nalezeno 19)*, ACM Turing Award *(značeno **TURING**, oceněných osob 57, v CiteSeer nalezeno 16, v DBLP nalezeno 12)*, ACM Fellows *(značeno **Fellows**, oceněných osob 809, v CiteSeer nalezeno 247, v DBLP nalezeno 192)* a ISI Highly Cited *(značeno **ISI**, oceněných osob 364, v CiteSeer nalezeno 146, v DBLP nalezeno 91)*.

Pouze autoři ocenění ACM SIGMOD E. F. Codd I. A. byli v databázích nalezeni manuálně. U zbylých ocenění byl použit postup, kdy jsme v databázích nalezli příjmení

oceněných autorů, z nichž jsme použili pouze ta, která se v databázi vyskytovala v méně jak 20 variacích a pro ně jsme nastálo určili jednoho zástupce, který byl v žebříčcích hodnocen nejlépe (pozn.: ve všech získaných žebříčcích se jednalo vždy o stejné osoby).

5 Výsledky

Všechny zmíněné postupy vyhodnocení jsme realizovali a z nich získané žebříčky autorů porovnali s udílenými oceněními, abychom zjistili, která varianta vyhodnocení je nejbližší kterému ocenění. Výsledek porovnání viz tab. 2 pro CiteSeer a tab. 3 pro DBLP.

V každém získaném žebříčku byli vždy nalezeni danou udílenou cennou ocenění autoři a jejich pořadí v žebříčku se sečetla, viz sloupec „součet“ (pozn.: pokud dva autoři skončili v žebříčku na stejné pozici, např. pozice 3-4, tak oba získali pořadí 3,5. Součty pořadí byly zaokrouhleny na celá čísla). Když v rámci udílené ceny seřadíme získané součty od nejmenšího (nejlepšího) k největšímu (nejhoršímu) a v tomto pořadí očíslujeme, tak získáme sloupec „p.“, tedy pořadí úspěšnosti jednotlivých variant vyhodnocení v rámci dané udílené ceny. V každém sloupci udíleného ocenění bylo navíc zvýrazněno 6 nejlepších (tučné černé písmo na šedém pozadí) a 6 nejhorších (bílé písmo na téměř černém pozadí) variant vyhodnocení.

Z obou tabulek (tab. 2 a tab. 3) je patrné, že se pořadí úspěšnosti jednotlivých variant vyhodnocení liší v závislosti na použité bibliografické databázi. Zatím, co u CiteSeer se úspěšnost jednotlivých variant liší pouze ve sloupci CODD a ve zbylých třech se velmi podobá, tak u DBLP nalezneme podobnost pouze ve sloupcích FELLOWS a ISI.

Další zajímavostí je, že nejúspěšnější varianty se pro jednotlivá ocenění a databáze liší, ale nejhorší varianty jsou v mnoha případech shodné a jsou jimi ohodnocení autorů získaná na základě hodnocení jejich publikací. Výjimku zde tvoří pouze sloupec CODD u CiteSeer a sloupec TURING u DBLP.

Jak již bylo řečeno, ***u CiteSeer*** se úspěšnost jednotlivých variant liší pouze ve sloupci CODD, kde tři nejlepší vyhodnocení byly získány ze sítí autorských citací, které mají váhy hran nastaveny na jedna a jsou vyhodnoceny pomocí upraveného vzorce PageRanku (viz vzorec (2)). Různé alternativy samocitací zde nehrály přílišnou roli. Za povšimnutí ale stojí čtvrtá nejlepší varianta vyhodnocení, kterou zde je aplikování upraveného vzorce PageRanku (viz vzorec (5)) na citační síť publikací bez samocitací a přenesení skóre PageRanku z publikací na autory tak, že každý z nich získá součet celých skóre svých publikací (viz vzorec (4)).

Úspěšnost jednotlivých variant vyhodnocení u zbylých třech ocenění se v CiteSeer liší minimálně. Nejlepší výsledky byly získány neupraveným vzorcem PageRanku (viz vzorec (1)) ze sítí autorských citací, které buďto neuvažují žádné samocitace, nebo váhy hran v těchto sítích jsou rovny jedné, popř. mají obě zmíněné vlastnosti.

U DBLP se úspěšnost jednotlivých variant vyhodnocení podobá pouze ve sloupcích FELLOWS a ISI, kde nejlepší výsledky poskytuje vyhodnocení sítí autorských citací upraveným vzorcem PageRanku (viz vzorec (2)). Tyto sítě opět buďto neuvažují žádné samocitace, nebo mají váhy všech hran v síti rovny jedné.

Ve sloupci CODD můžeme vidět, že nejlepší varianty pro toto udílené ocenění pracovaly se sítí autorských citací, kde hrany vyjadřují, kolikrát autor citoval jiného autora (znač. **N**), přičemž téměř nezáleželo na tom, jakým způsobem zacházíme se samocitacemi a

zda použijeme upravený nebo neupravený vzorec PageRanku (pozn.: varianta pracující se všemi samocitacemi a upraveným vzorcem PageRanku zde byla nejlepší).

Sloupec TURING u DBLP byl z obou databází jediný, kde všechny nejlepší varianty vyhodnocení byly získány přenesením skóre PageRanku z publikací na autory. Přičemž nejlepších výsledků dosáhly varianty přenosu, které rovnoměrně dělily skóre publikace mezi její autory (viz vzorec (3)). Použití neupraveného, či upraveného vzorce PageRanku nemělo přílišný vliv na hodnocení, ale nejlepší výsledek poskytl vzorec neupravený (tedy vzorec (1)).

Síť	**typ sítě**	**typ vah**	**vzorec**	**CiteSeer**							
				FELLOWS		**TURING**		**CODD**		**ISI**	
				součet	p.	součet	p.	součet	p.	součet	p.
Autoři	**ŽÁDNÉ**	**1**	Neupr.	**10658192**	**1**	**834687**	**1**	87976	16	**6144626**	**1**
			Upr.1	**11733294**	**6**	**950528**	**6**	**70741**	**3**	6676047	7
		1/N	Neupr.	**11120326**	**3**	**874238**	**2**	110065	26	**6452255**	**4**
			Upr.1	12847770	15	1027517	14	83104	14	7296809	15
		N	Neupr.	**10921269**	**2**	**874471**	**3**	103524	25	**6259453**	**2**
			Upr.1	12359833	13	1017987	12	79391	9	6956451	12
	SPOLUAUT	**1**	Neupr.	**11469941**	**4**	**938763**	**5**	81735	12	**6426664**	**3**
			Upr.1	12015292	8	1021031	13	**66852**	**1**	6780031	8
		1/N	Neupr.	12049050	9	977371	7	97157	22	6813669	10
			Upr.1	12980099	17	1078353	17	78668	8	7484989	17
		N	Neupr.	11914793	7	982140	8	88955	18	**6657198**	**6**
			Upr.1	12708448	14	1076750	15	**72922**	**5**	7176417	14
	VŠE	**1**	Neupr.	**11620341**	**5**	**935368**	**4**	84168	15	**6527415**	**5**
			Upr.1	12119980	10	1015760	11	**68316**	**2**	6853879	11
		1/N	Neupr.	12290050	12	991516	10	99001	23	6968269	13
			Upr.1	13230123	18	1088803	19	79872	11	7647983	18
		N	Neupr.	12122379	11	985738	9	90804	19	6788121	9
			Upr.1	12897011	16	1078023	16	**73971**	**6**	7323504	16
Publikace	**ŽÁDNÉ**	**části**	Neupr.	17926089	26	1229962	22	95424	21	10428867	26
			Upr.2	17134450	24	1228119	21	88217	17	9991005	24
		celé	Neupr.	14047488	20	1080128	18	74034	7	8241516	19
			Upr.2	15300249	22	1238358	23	**72026**	**4**	8980913	21
	VŠE	**části**	Neupr.	17328218	25	1265683	25	103011	24	10055987	25
			Upr.2	16353127	23	1240312	24	93938	20	9563706	23
		celé	Neupr.	13944487	19	1129859	20	82537	13	8272036	20
			Upr.2	15007515	21	1276666	26	79505	10	9008220	22

Tab. 2. Aplikování PageRanku na citační sítě vzniklé ze CiteSeer.

síť	typ sítě	typ vah	vzorec	DBLP							
				FELLOWS		TURING		CODD		ISI	
				součet	p.	součet	p.	součet	p.	součet	p.
Autoři	ŽÁDNÉ	1	Neupr.	729899	10	40672	21	897	22	365682	12
			Upr.1	**642752**	**1**	28586	13	778	13	**311325**	**1**
		1/N	Neupr.	**696265**	**5**	36711	18	781	14	345308	7
			Upr.1	**684210**	**4**	26363	9	740	11	**333845**	**4**
		N	Neupr.	722662	8	41100	22	**658**	**6**	365580	11
			Upr.1	**699794**	**6**	27718	12	**613**	**4**	**343010**	**6**
	SPOLUAUT	1	Neupr.	805467	15	41743	23	887	20	403208	15
			Upr.1	**678709**	**2**	28613	14	775	12	**323612**	**2**
		1/N	Neupr.	767966	13	36889	19	715	10	379210	13
			Upr.1	710798	7	27123	10	685	8	**337727**	**5**
		N	Neupr.	807944	16	42417	24	**630**	**5**	407987	16
			Upr.1	737243	11	28910	15	**580**	**2**	353612	9
	VŠE	1	Neupr.	815327	17	42512	25	896	21	408281	17
			Upr.1	**682244**	**3**	29008	16	782	15	**325416**	**3**
		1/N	Neupr.	779763	14	38095	20	708	9	385298	14
			Upr.1	728208	9	27685	11	677	7	353529	8
		N	Neupr.	819845	18	43561	26	**612**	**3**	414114	18
			Upr.1	743486	12	29433	17	**570**	**1**	356039	10
Publikace	ŽÁDNÉ	části	Neupr.	1254475	25	**19560**	**1**	1548	26	639001	26
			Upr.2	1172179	22	**19753**	**3**	1495	23	586390	21
		celé	Neupr.	1075760	19	**24159**	**5**	854	19	525481	20
			Upr.2	1237086	23	25923	7	850	18	590908	22
	VŠE	části	Neupr.	1255171	26	**19718**	**2**	1527	25	634750	25
			Upr.2	1167223	21	**19949**	**4**	1499	24	600034	24
		celé	Neupr.	1077594	20	**24395**	**6**	840	17	519665	19
			Upr.2	1248008	24	26178	8	819	16	592794	23

Tab. 3. Aplikování PageRanku na citační sítě vzniklé z DBLP.

Za věrohodnější, ve vztahu k oběma kolekcím, a tedy i za více směrodatné, můžeme považovat porovnání s oceněními ACM Fellows a ISI Highly Cited, která mají značně větší počet oceněných osob, než zbylá dvě ocenění.

Pokud shrneme získané výsledky, můžeme říci, že v subjektivním hodnocení autorů (tedy oceněních) se samocitace příliš neuplatní. Tomu odpovídají sítě autorských citací, které neobsahují žádné samocitace. Z výsledků dále plyne, že významnost autora závisí více na tom, kolik jiných autorů daného autora citovalo (sítě s váhami značenými 1), než kolikrát ho citovali (sítě s váhami značenými N), a že uvažování počtu spoluautorů při

daných hlediscích nehraje přílišnou roli při hodnocení autora (sítě s váhami značenými 1 poskytovaly lepší výsledky než sítě s váhami značenými 1/N), oproti např. RIV (Rejstřík informací o výsledcích), který dělí hodnotu publikace mezi její autory. Také nemá smysl při hodnocení autorů vycházet přímo z hodnocení publikací, jelikož výsledky vyhodnocení pracujících s citačními sítěmi publikací byly obvykle nejhorší.

V CiteSeer bylo nejlepších výsledků dosaženo vzorcem (1), kdežto v DBLP vzorcem (2). To lze přisoudit skutečnosti, že DBLP obsahuje málo citačních vazeb a tudíž se v hodnocení více projevil vliv první části vzorce (2) zohledňující počet publikací autora.

Pro obě bibliografické databáze jsme navíc provedli přepočítání vzorce (1) pro síť autorských citací bez samocitací s váhou hran 1 a damping faktorem 0,9. Získané žebříčky jsme porovnali s kompletními žebříčky autorů zapůjčenými od Dalibora Fialy (viz [2] a [3]), které získal pomocí své implementace „standard PageRank“. Spearmanův koef. korel. byl cca 0,92 pro obě databáze a Kendallův koef. korel. byl 0,89 pro CiteSeer a 0,81 pro DBLP, přičemž rozdíl lze přisoudit např. odlišnému způsobu zacházení s uzly bez výstupní hrany.

6 Závěr

V článku jsme představili několik variant vyhodnocení citačních sítí algoritmem PageRank a získané žebříčky autorů porovnali s významnými institucemi udílenými oceněními. Ukázalo se, že úspěšnost jednotlivých variant vyhodnocení závisí na zvolené bibliografické databázi, z níž je citační síť vytvořena, a také na udíleném ocenění, se kterým vypočtený žebříček autorů porovnáme. Obecně dobré výsledky byly v tomto případě (data, parametry) obvykle získány ze sítí autorských citací, které neuvažují žádný druh samocitací. Metody přenosu PageRanku z publikací na autory se tedy neosvědčily (vyjma porovnání ACM Turing Award s výsledky získanými z DBLP). Význam ostatních úprav sítě, či použitého vzorce PageRanku na úspěšnost dané varianty vyhodnocení se již značně lišil u každého udíleného ocenění.

Na základě zjištěných poznatků si myslíme, že využitím výsledků citační analýzy lze určit rozdíl ve významnosti dvou osob dle rozdílu jejich hodnot PageRanku (popř. pořadí ve výsledném žebříčku), k čemuž by mohlo být přihlédnuto např. při přidělování grantů, či při hledání osob na vedoucí pozice ve výzkumných institucích. Lze podotknout, že o grant se vždy žádá v určité vědní oblasti (např.: informační vědy), pro kterou buďto existuje odpovídající bibliografická databáze, nebo lze databázi vytvořit z časopisů a konferenčních sborníků zabývajících se danou oblastí vědy.

V další práci bychom rádi porovnali naše výsledky s výsledky získanými pomocí Centrality Measure, h-indexu a případně s žebříčkem získaným na základě hodnocení RIV. Chtěli bychom také zopakovat v tomto článku nastíněné postupy vyhodnocení autorů pro bibliografickou databázi ISI Web of Science.

Literatura

1. Ding, Y.: Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology* vol. 62 (2011) 236-245.

2. Fiala, D.: Mining citation information from CiteSeer data. *Scientometrics* vol. 86 (2010) 553-562.

3. Fiala, D.: *Web mining methods for the detection of authoritative sources*. Ph.D. dissertation, Department of Computer Science and Engineering, University of West Bohemia, Pilsen, Czech Republic, 2007.

4. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring Web Communities from Link Topology. In: *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh (1998) 225–234.

5. Heller, P.: *Vyhodnocování autoritativních zdrojů na bázi PageRanku*. Diplomová práce, Katedra informatiky a výpočetní techniky, Západočeská univerzita v Plzni, Plzeň, Česká Republika, 2010.

6. Langville, A.N., Meyer, C.D.: *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton, New Jersey, 2006.

7. Nykl, M.: *Vyhodnocování informačních sítí*. Diplomová práce, Katedra informatiky a výpočetní techniky, Západočeská univerzita v Plzni, Plzeň, Česká Republika, 2011.

8. Page, L., Brin, S., Motwani, R., Winograd, T.: *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.

9. Yan, E., Ding, Y., Sugimoto, C.R.: P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology* vol. 62 (2011) 467-477.

10. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology* vol. 62 (2009) 2107-2118.

11. Zhao, D.: Going beyond counting first authors in author co-citation analysis. In: *Proceedings of the American Society for Information Science and Technology* vol. 42 (2005).

Annotation:

PageRank variants for citation analysis

The paper will introduce the issue of evaluation of citation networks, which aims to determine the significance of the authors. We show the evaluation of citation networks using PageRank algorithm with respect to co-authorship and compare the calculated orders of authors with the most famous annual awards such as the ACM A.M. Turing Award, ACM SIGMOD E.F. Codd Innovation Award, ACM Fellow and ISI Highly Cited. Such analysis of citation networks can be used for allocation of grants, or searching of persons for leading positions in research institutions.
In this paper we will go through the whole procedure of obtaining citation networks from bibliographic data, determining weights of edges in these networks, PageRank evaluation and comparing results of these rankings with the given awards.