# PageRank a vyhodnocování citačních sítí s ohledem na spoluautorství

**Michal Nykl**[1]

## 1 Algoritmus PageRank a sociální sítě autorů

Vyhodnocování sociálních publikačních sítí autorů vědeckých článků, za účelem vyhledání „kvalitních“ vědců, se stalo nedílnou součástí např. přidělování grantů, či hledání osob na vedoucí pozice ve výzkumných zařízeních.

Jednou z možností, jak vyhodnotit autory v závislosti na jejich publikační činnosti a kvalitě jejich publikací (výzkumu), je použít algoritmus PageRank (viz Langville and Meyer (2006)) na sociální síť, kterou daní autoři tvoří, a následně autory seřadit dle hodnot, které jim algoritmus vypočítá. V takto vytvořené sociální síti každého autora zastupuje jeden uzel a (orientované) hrany, včetně vah, určitým způsobem zohledňují vzájemné citování autorů, či jejich spoluautorství.

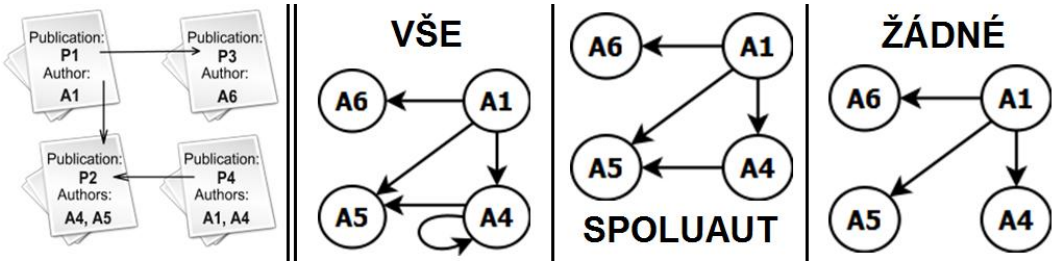
## 2 Sítě autorských citací, spoluautorů a společně-citovaných autorů

Algoritmus PageRank byl již v minulosti několikrát uplatněn na síť autorských citací, na síť spoluautorů, či na síť společně-citovaných autorů, kde hrany vyjadřují, že autoři byli vzájemně citováni jiným autorem (v jedné jeho publikaci). Váha hrany vždy určitým způsobem odráží míru dané skutečnosti. Relevantní práce např. viz Nykl (2011), Zhao (2005), či Yan a Ding (2011).

Dále v textu se zaměříme na to, jakým způsobem lze vytvářet sítě autorských citací tak, aby hrany a váhy vznikaly s určitým ohledem na spoluautorství, a jaký vzorec PageRanku použít při jejich vyhodnocování. Jednou z dalších užitých (v článku nezmíněných) alternativ vyhodnocování je určení hodnot PageRanku všech publikací (z citační sítě publikací) a předání hodnoty PageRanku publikace jejím autorům.

## 3 Tvorba hran sítě autorských citací s ohledem na samocitace a spoluautorství

Jedním z aspektů při vytváření sítě autorských citací je, jak moc jsme ochotni připustit, aby „kvalita“ 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 (znač. VŠE), b) odstraníme citace mezi publikacemi se stejným autorem (znač. ŽÁDNÉ), c) pokud autor cituje svou publikaci, pak cituje pouze své spoluautory v této publikaci, ale necituje sebe (znač. SPOLUAUT), příklad viz obr. 1. (Pozn.: v případě VŠE můžeme posléze např. penalizovat váhu hran představujících samocitace.)




**Obrázek 1:** Způsoby zohlednění samocitací při tvorbě sítě autorských citací (nalevo citační síť publikací).

[1] student doktorského studijního programu Inženýrská informatika, obor Informatika a výpočetní technika, specializace Modely a metody extrakce informací z webu, e-mail: nyklm@kiv.zcu.cz

## 4 Určení vah hran v síti autorských citací

Váhu hran v síti autorských citací lze, s ohledem na spoluautorství, určit mnoha způsoby, např. viz Fiala (2007). My zde nyní porovnáme způsoby tři a to: a) každá hrana v citační síti autorů má váhu jedna, b) váha hrany vyjadřuje, kolikrát autor citoval publikace daného autora, c) pokud autor citoval publikaci se dvěma autory, pak z jeho uzlu povede na uzel každého z těchto autorů hrana s váhou ½ (pozn.: souhlasné hrany mezi dvěma autory se spojí a jejich váhy se sečtou).

## 5 Zvolený vzorec PageRanku

Pro vyhodnocování vzniklých citačních sítí s rozličným ohodnocením hran byl použit iterační vzorec PageRanku, viz (1), kde $P_x(A)$ je hodnota PageRanku uzlu $A$ v iteraci $x$, $d$ je damping faktor (obvykle 0,85), $|V|$ 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.

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

## 6 Závěr

V článku nastíněné vytváření sítí autorských citací, přiřazování vah hranám v těchto sítích a následné vyhodnocení sítí algoritmem PageRank bylo použito k vyhodnocení sítí autorských citací, které vznikly zmíněnými postupy z bibliografických databází CiteSeer a DBLP.

Získané výsledky jsme následně porovnali se seznamy oceněných osob, které jsme vytvořili z každoročně udílené ceny ACM SIGMOD E.F.Codd Innovation Award, z ceny ACM A.M. Turing Award, z držitelů ocenění ACM Fellows a z hodně citovaných autorů ISI Highly Cited Researchers. Cílem bylo určit, které výše uvedené úpravy sítě poskytují, při vyhodnocení algoritmem PageRank, pořadí autorů s oceněnými osobami na co nejlepších pozicích. Nejzajímavější výsledky budou ukázány při prezentaci.

V další práci aplikujeme zmíněné postupy vyhodnocování na sítě vzniklé z bibliografické databáze ISI Web of Science. Chtěli bychom také dané sítě vyhodnotit pomocí dalších metrik, jako např. Centrality Measure nebo HITS a pokusit se pomocí dostupných prostředků analyzovat sociální síť www.lide.cz.

## Literatura


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

Nykl, M., 2011. *Vyhodnocování informačních sítí*. Ing. thesis, University of West Bohemia, Pilsen, CR.

Zhao, D., 2005. Going Beyond Counting First Authors in Author Co-citation Analysis. *Proceedings of the American Society for Information Science and Technology*, Vol. 42.

Yan, E., and Ding, Y., 2011. Discovering author impact: A PageRank perspective. *Information Processing and Management*, Vol. 47. pp 125-134.

Fiala, D., 2007. *Web mining methods for the detection of authoritative sources*. Ph.D. thesis, University of West Bohemia, Pilsen, CR.