VYHODNOCOVÁNÍ INFORMAČNÍCH SÍTÍ NA BÁZI PAGERANKU

Michal NYKL[1]

1 KLASICKÝ ALGORITMUS PAGERANK

Algoritmus PageRank patří mezi úspěšně používané algoritmy umožňující řazení uzlů sítě na základě vstupních hran a významnosti uzlů, ze kterých tyto hrany vedou. Používán je např. v internetových vyhledavačích (Google.com), kde se vedle fulltextového vyhledávání podílí na řazení výsledků hledání, a již několikrát byl též úspěšně uplatněn při vyhodnocování informačních sítí.

Klasický algoritmus PageRank, tak jak jej ve své původní práci uvádějí Page et al. (1998), byl navržen pro vyhodnocování webové sítě, kde uzly představují webové stránky a hrany reprezentují hypertextové odkazy mezi nimi, přičemž z uzlu může vést do jiného uzlu maximálně jedna hrana. Hrana pouze říká, že z první webové stránky na druhou existuje hypertextový odkaz, ale neříká, kolik odkazů, viz vzorec (1), kde $P_{x+1}(A)$ představuje hodnotu PageRanku uzlu A v iteraci x+1, d je damping faktor (udává pravděpodobnost s jakou se „surfař" dostane na uzel pomocí hrany vedoucí z jiného uzlu - druhá část vzorce (1), či přijde na uzel zcela náhodně - první část vzorce (1)). $|V|$ je počet uzlů sítě, U je množina uzlů, ze kterých vede hrana na uzel A, $P_x(u)$ je PageRank uzlu u ($u \in U$) v iteraci x a N_u je počet uzlů, na které vede hrana z uzlu u.

$$P_{x+1}(A) = \frac{(1-d)}{|V|} + d \sum_{u \in U} \frac{P_x(u)}{N_u} \quad (1)$$

Před vyhodnocováním sítě klasickým algoritmem PageRank je potřeba ze sítě odstranit hrany vedoucí ze/do sítě a případně odstranit hrany vystupující z téhož uzlu, do něhož vedou (tj. v síti zůstanou pouze hrany, které vedou mezi rozdílnými uzly téže sítě). Výpočet PageRanku je iterativní, přičemž stačí, když každý uzel získá počáteční hodnotu PageRanku rovnu 1/|V|, a končí dosažením ukončovacího kritéria (např. počet iterací nebo ustálení výsledků – hodnot PageRanků, či výsledného pořadí uzlů).

2 OŠETŘENÍ UZLŮ BEZ VÝSTUPNÍCH HRAN

Hodnota PageRanku uzlu sítě vyjadřuje pravděpodobnost, s jakou daný uzel navštíví „surfař", tj. člověk, který se v síti pohybuje (prochází uzly dle hran, či zcela náhodně – dáno damping faktorem), a hodnota součtu PageRanku všech uzlů sítě by tedy měla být rovna jedné (100%).

S tímto souvisí jeden z problémů klasického algoritmu PageRank a to ubývání hodnoty součtu PageRanku všech uzlů sítě vlivem uzlů bez výstupních hran. Tento problém lze řešit např. odstraněním hran vedoucích do uzlů bez výstupních hran (tzv. dangling links), či odstraněním uzlů bez výstupních hran ze sítě, což ovšem bývá rekurzivní a může vést k odstranění celé sítě. Proto lepším řešením (připustíme-li, že z uzlu bez výstupní hrany musí

[1] Michal Nykl, student navazujícího (inženýrského) studijního oboru Softwarové inženýrství na Západočeské univerzitě v Plzni (ZČU), Fakulta aplikovaných věd (FAV), Katedra informatiky a výpočetní techniky (KIV).
e-mail: nyklm@students.zcu.cz Vedoucí diplomové práce: Doc. Ing. Karel Ježek, CSc.

„surfař“ přejít na libovolný jiný uzel, aby mohl pokračovat v procházení sítě) je provázání uzlů bez výstupních hran se všemi uzly sítě, které může být provedeno „fyzicky“ (tj. v síti vzniknou nové hrany) nebo upravením vzorce PageRanku, např. viz vzorec (2) – vzorec algoritmu PageRank ošetřující uzly bez výstupních hran, kde ***D*** představuje množinu uzlů bez výstupních hran a suma v čitateli posledního zlomku vzorce (2) vyjadřuje součet hodnot PageRanku uzlů z množiny ***D***, ze které je vynechán prvek ***A***, pokud do ní patří.

$$P_{x+1}(A) = \frac{(1-d)}{|V|} + d * \left(\sum_{u \in U} \frac{P_x(u)}{N_u} + \frac{\sum_{s \in D, s \neq A} P_x(s)}{|V|-1}\right) \quad (2)$$

3 UVAŽOVÁNÍ MNOŽSTVÍ HRAN MEZI DVOJICEMI UZLŮ

Neuvažování množství hran mezi dvojicemi uzlů je typické pro webovou síť (či síť citací mezi publikacemi), kde není důvod, aby z jednoho uzlu vedlo více hran do uzlu jiného. Situace se ovšem změní např. u sítě autorských citací, kde autor může citovat více publikací jiného autora, tj. v síti povede z původního autora na jiného autora více hran.

Vzorec PageRanku lze upravit tak, aby množství hran uvažoval jen v některé své části (první nebo druhé), či v obou svých částech vzorce, viz vzorec (3), kde N_c je počet všech hran v síti, N_{Aci} je počet vstupních hran uzlu ***A***, N_{uco} je počet výstupních hran uzlu ***u*** a ***CC(u,A)*** je počet hran vedoucích z uzlu ***u*** do uzlu ***A***. Vzorec (3) opět neošetřuje uzly bez výstupních hran (v rámci práce vznikly i vzorce PageRanku uvažující množství hran (v první, druhé, či obou částech vzorce) a ošetřující uzly bez výstupních hran).

$$P_{x+1}(A) = (1-d)\frac{N_{Aci}}{N_c} + d\sum_{u \in U} \frac{P_x(u) * CC(u,A)}{N_{uco}} \quad (3)$$

4 VYHODNOCENÍ INFORMAČNÍCH SÍTÍ

Všechny zde zmíněné algoritmy byly vzájemně porovnány (pomocí koeficientů korelace) a posléze aplikovány na sítě autorských citací, sítě citací mezi afiliacemi, sítě spoluautorství a sítě spolupráce afiliací vytvořené z bibliografických databází DBLP (2004) a CiteSeer (2005), za účelem vyhodnocení významnosti autorů a afiliací obsažených v těchto databázích. Algoritmy vytvořená pořadí autorů (sestupné řazení dle hodnoty PageRanku) byla též srovnána s vítězi každoročně udílené ceny E. F. Codd Innovation Award.

5 ZÁVĚR

Přínosem práce je modifikace klasického algoritmu PageRank ošetřující uzly bez výstupních hran a modifikace algoritmu PageRanku uvažující množství hran (v první, druhé nebo obou částech vzorce) a navíc případně též ošetřující uzly bez výstupních hran. V rámci práce vznikla knihovna umožňující vytváření citačních sítí a sítí spolupráce (publikací, autorů, afiliací) z dat získaných z databází DBLP (2004) a CiteSeer (2005), která umožňuje aplikovat všechny v práci zmíněné algoritmy na libovolnou z těchto sítí a porovnat výsledná pořadí (sestupné řazení dle hodnoty PageRanku) pomocí Spearmanova koeficientu korelace. Navíc vznikla knihovna a aplikace umožňující výpočet Spearmanova a Kendallova koeficientu korelace.

LITERATURA

Page, L., and Brin, S., and Motwani, R., and Winograd, T., 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab, California.