Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra infromatiky a výpočetní techniky

Bakalářská práce

Měření významnosti autorů v citační síti

Plzeň, 2013

Tomáš Maršálek

Abstrakt

Prvky sociální sítě, které nemají žádné apriorní ohodnocení významnosti, jsou různě významné pouze na základě vztahů s okolními prvky. V této práci byly prozkoumány a implementovány známé metody centrality a bibliografické metody měřící významnost prvků v sociální nebo citační síti. Výsledky aplikování metod na volně dostupné citační databáze ukázaly vysokou podobnost jednotlivých metod a rovněž shodu nejvýznamnějších autorů dle těchto metod se známými oceněními v oblasti informatiky a informační vědy (Turing Award, Codd, ACM Fellows, ISI Highly Cited). Bylo zjištěno, že některé implementované metody jsou i přes použití nejrychlejších algoritmů výpočetně příliš náročné vzhledem k velikosti citačních sítí, vzniklých z těchto citačních databází. Dále bylo empiricky potvrzeno, že implementované metody měří významnost, byť může mít více interpretací.

Obsah

KAPITOLA 1

Úvod

KAPITOLA 2

Sociální a citační sítě

2.1 Sociální sítě

Myšlenka sociální sítě existovala dlouho předtím, než je pod tímto termínem začali lidé rozpoznávat. Jedná se o komplexní struktury vztahů mezi členy sociálních uspořádání na všech úrovních - od osobních až po mezinárodní vztahy mezi organizacemi.

Nejčastěji se ale setkáme se sociální sítí jako strukturou tvořenou lidmi, kteří jsou svázáni nějakým sociálním vztahem. Nejčastěji, zejména v poslední době s rozmachem populárních webových sociálních sítí (MySpace, Facebook, G+, Lidé), jím bývá přátelství.

2.2 Analýza sociálních sítí

2.2.1 Bezškálové sítě

Více než 40 let byly považovány všechny komplexní sítě za naprosto náhodné. Paul Erdős a Alfréd Rényi v roce 1959 navrhli modelování komunikačních sítí a sítí, které se vyskytují v přírodních vědách, spojením uzlů náhodnými hranami. Tento jednoduchý způsob způsobí rozložení stupňů vrcholů podle Poissonova rozdělení s charakteristickou křivkou připomínající zvon - většina

uzlů má zhruba stejný stupeň. V roce 1998 bylo na univerzitě v Notre Dame (Barabási a kolegové) provedeno mapování sítě World Wide Web s očekáváním, že výsledkem bude náhodná síť. Přestože byl zmapován pouze zlomek celé sítě, výsledkem bylo přes všechna očekávání, zcela jiné rozdělení stupňů - mocninné. Přes 80% uzlů mělo méně než čtyři spojení, ale méně než 0.01% uzlů mělo více než tisíc spojení. Sítě, které se řídí mocninným rozdělením, nazvali Barabási a jeho kolegové bezškálovými sítěmi (scale-free network). Rozpoznání tohoto jevu vedlo k lepšímu porozumění šíření virů a epidemií nebo proč některé sítě fungují takřka beze změny i přes poruchu většiny jejich uzlů. Sociální a citační sítě se řadí do kategorie bezškálových sítí. Například autor vědecké literatury, jehož dílo je v dané oblasti známé, má velkou šanci, že bude citován dalšími autory, především těmi novými. Stejně tak osoba v sociální síti s velkým počtem přátel má velikou šanci, že bude představen novým lidem a rozšíří si tak svůj kruh přátel ještě více. Tomuto jevu v bezškálových sítí se říká „bohatší se stává bohatším“.

2.3 Citační sítě

Citační sítě jsou podobné sociálním sítím, pouze místo uzlů, které představují osoby, v citační síti se jedná o publikace nebo autory těchto publikací. Pokud je uzlem publikace, pak hrany této sítě symbolizují citaci publikace jinou publikací. V druhém případě uvažujeme síť, kde uzly reprezentují autory knih, vědeckých článků, vědecké literatury a dalších publikací. Prvnímu typu říkáme síť publikací, druhému síť autorů.

2.3.1 Síť publikací

Uvažujeme-li první případ, kde uzly reprezentují publikace a hrany přímo citace mezi těmito publikacemi, jedná se o síť publikací. Tedy pokud publikace A odkazuje na publikaci B, pak budou existovat stejnojmenné uzly A a B a hrana mezi těmito uzly může mít dvě různé orientace podle svého uplatnění. Směr od citující publikace k citované (v našem příkladě od A do B) bude mít hrana, kterou označíme jako výstupní pro uzel A a vstupní pro uzel B. Výstupní hrana laicky řečeno označuje vztah "cituji", kdežto vstupní hrana znamená "jsem citován".

2.3.2 Síť autorů

Druhým případem citační sítě je síť autorů. Zde je uzel reprezentací autora a hrany spojují autory mezi sebou. Ve většině případech máme k dispozici data ve formátu, který přímo odpovídá síti publikací, tzn. pro jednu publikaci známe seznam jejích autorů a odkazů na další publikace. Síť autorů lze získat transformací sítě publikací tak, že každou hranu z původní sítě publikací přiřadíme každému z autorů této publikace a duplikujeme ji pro každého z autorů citované publikace. Celkově vznikne nm nových hran, pokud odkazovaná publikace obsahuje n autorů a odkazující m autorů. Stejně jako v síti publikací, i zde uvažujeme dvě opačné orientace hrany se stejnou interpretací, tedy "cituji"a "jsem citován".

V síti autorů má pro naše účely smysl uvažovat ohodnocení hran. Existuje více způsobů, jak přiřadit ohodnocení (váhy) jednotlivým hranám, ale nejjednodušším způsobem, který je použitý i v implementaci knihovny, je prosté přiřazení počtu publikací, jejichž autorem nebo spoluautorem je daný autor A, které odkazují na publikace, jejichž autorem je autor B. Srozumitelnější popis poskytne obrázek:

Druhým způsobem ohodnocení hran, který rovněž využívá implementovaná knihovna pro některé metody, je převrácená hodnota prvního způsobu ohodnocení. Důvodem je přímá souvislost mezi vahou hrany a vzdáleností mezi uzly. V prvním případě, kdy silnější pouto mezi autory vyjadřuje vyšší ohodnocení hrany, v druhém případě je naopak nižší váha vyjádřením silnějšího vztahu, jelikož jsou si uzly blíže. Tento způsob je používán pro algoritmy, které pracují na myšlence nejkratších cest mezi uzly.

2.3.3 Vážené citační sítě

V definici grafu nebo sítě $G = (V, E)$ je množina hran E soubor dvojic, které označují koncové uzly hrany, neboli jejich spojení. Samotné spojení je jediná informace, kterou množina hran nese. Chceme-li zaznamenat nějakou další informaci, která je spojená se spojením dvou uzlů, namísto hrany jako dvojice koncových uzlů nadefinujeme hranu jako n-tici, kde první dvě hodnoty jsou koncové uzly a zbylé hodnoty nesou libovolnou informaci. Ve většině případů si vystačíme s jednou dodatečnou informací a nazýváme ji váha hrany.

Při zavedení vah máme například možnost používat síť jako multigraf, tedy graf, u kterého je povoleno více spojení mezi dvěma stejnými uzly. Počet stejných hran pak pouze zaznamenáme celočíselnou hodnotou ve váze hrany.

Například síť world wide web tvořená webovými stránkami je příkladem multigrafu, protože je povoleno z jedné stránky odkazovat na jinou na více místech. Při analýze takových sítí využijeme právě vah hran a počet hypertextových odkazů mezi dvěma stránkami zaznamenáme vyšším ohodnocením hrany. V tomhle případě znamená vyšší váha silnější pouto mezi uzly.

Jiným případem může být například síť kde sledujeme města a dopravní spojení mezi nimi. V tomhle případě nás může zajímat vzdálenost nebo časová náročnost na dopravu mezi dvěma městy, které budou znamenat silnější pouto pokud budou mít naopak menší váhu. Hledáme totiž nejkratší a nejrychlejší spojení.

Pro citační sítě můžeme uvažovat ohodnocení hran obojího typu. Například mezi dvěma autory může být silnější vztah, pokud se citují ve více publikacích. Pokud citační síť analyzujeme metodami, které jsou založené na myšlence hledání nejkratších cest i v této síti, která nemá v podstatě žádný pojem vzdálenosti, použijeme druhý typ ohodnocení - menší váha, silnější pouto.

2.3.4 Orientované a neorientované sítě

Obecně můžeme uvažovat grafy s hranami s orientací či bez orientace. V obou případech se stále jedná o množinu (V, E), pouze pro orientovaný graf je množina hran množinou uspořádaných dvojic oproti množině neuspořádaných dvojic u neorientovaného grafu.

Hrany se uzlu v případě orientovaného grafu liší z pohledu jednoho uzlu. Pokud hrana vychází z tohoto uzlu, nazveme ji výstupní hrana, v opačném případě se bude jednat o vstupní hranu.

V případě sociálních sítí nejčastěji uvažujeme sítě bez orientace, protože nejčastěji modelovaný vztah přítel-přítel je ekvivalentní z pohledu obou koncových uzlů. Pro citační síť jsou na místě orientované hrany, protože vztahy autor odkazujícího na jiného autora nebo publikace citující jinou publikaci mají očividně jinou interpretaci z pohledu koncových uzlů. Buďto se jedná o citovaného nebo citujícího autora či publikaci.

2.3.5 Souvislost a komponenty grafu

Pro neorientovaný graf je komponenta maximálně souvislý podgraf. Jinak řečeno komponenta je podgraf takový, že všechny jeho vrcholy jsou spojeny nějakou cestou. Komponentou ji i samotný vrchol.

Všechny komponenty grafu najdeme pomocí jednoduchých algoritmů prohledávání do šířky nebo do hloubky. Spuštění prohledávání najde celou komponentu, ve které se výchozí vrchol nachází. Spustíme-li prohledávání ze všech vrcholů, najdeme všechny komponenty.

Slabě souvislý orientovaný graf znamená, že neorientovaný graf, který by vznikl nahrazením orientovaných hran neorientovanými (symetrizace grafu), by byl souvislý.

Pro zachování vlastnosti souvislosti, že všechny vrcholy jsou spojené nějakou cestou, pro orientovaný graf musíme uvažovat silně souvislý graf nebo podgraf. Definice zůstává stejná jako u slabě spojitých komponent, ale protože hrany nejsou oboustranné, mezi dvěma spojenými vrcholy ne vždy existuje cesta oběma směry. U neorientovaného grafu můžeme souvislost vyjádřit tak, že pro každé dva uzly u a v existuje cesta z u do v. Protože jsou hrany symetrické, pak automaticky existuje i cesta z v do u. U orientovaného grafu musíme druhou podmínku explicitně dodat: graf je silně souvislý, pokud pro každé dva vrcholy u a v existuje cesta z u do v i z v do u.

Silně spojité komponenty nenajdeme pouhým prohledáním do šířky nebo do hloubky, ale použijeme sofistikovanější algoritmy (Kosarajův, Tarjanův, ...), které ale vycházejí z prohledávání do hloubky.

2.3.6 Klika v grafu

Klika (clique) grafu je úplný podgraf. To znamená, že všechny vrcholy kliky jsou spojeny přímo hranou.

V sociologii slovo klika souvisí se skupinou lidí, kteří jsou na sebe vázáni více než na jiné lidi v tomtéž prostředí. Klika je silněji spojená skupina lidí než sociální kruh.

2.4 Analýza citačních sítí

2.5 Citační databáze

Citační databáze poskytují možnost vyhledávání bibliografických citací. Většina z dnešních citačních databází se zaměřuje na jeden obor. Full-textové databáze poskytují kompletní text publikací, které indexují (cite http://library.amnh.org/research-tools/citation-full-text-databases).

2.5.1 DBLP

DBLP [?] je webová bibliografická databáze hostovaná na univerzitě Trier. Od 80. let indexovala literaturu z oblasti databází a logického programovaní, ale postupně se její zaměření zobecnilo a nyní je bibliografickou databází obecně pro obor informatiky. V roce 2012 obsahovala více než 2,1 milionu článků. Metody implementované v této práci jsou aplikovány na verzi z roku 2004.

Charakteristika

Struktura sítě

Rozďělení vah

2.5.2 CiteSeer

CiteSeer (nyní CiteSeer^X) [?] je považován za první automatizovaný systém shromažďování publikací a autonomní indexace citací v nich obsažených. Publikace jsou zejména z oboru informatiky a informační vědy. V dnešní době obsahuje přes dva miliony dokumentů s téměř dvěma miliony autorů a čtyřiceti miliony citací. Zde používáme verzi z roku 2005.

2.6 Ocenění významných autorů

2.6.1 ACM A.M. Turing Award

ACM A.M. Turing Award je ocenění ročně udělované skupinou ACM (Association for Computing Machinery) jedincům vybraným pro kontribuce technického ducha do výpočetního světa. [?].

Turingova cena je brána jako nejvyšší vyznamenání v informatice a je lidově nazývána Nobelovou cenou pro informatiku [?, p. 317].

2.6.2 ACM SIGMOD Edgar F. Codd Innovations Award

ACM SIGMOD Edgar F. Codd Innovations Award je ohodnocení životního díla skupinou ACM SIGMOD (Special Interest Group on Management of Data) za inovativní a vysoce ceněné kontribuce k rozvoji, porozumění a použití databázových systémů a databází [?].

2.6.3 ACM Fellows

„The ACM Fellows Program“ byl založen v roce 1993, aby našel a ocenil vynikající členy ACM za jejich dílo v informatice a informační vědě a pro jejich významné kontribuce pro účel ACM. Členové ACM Fellows slouží jako význačtí kolegové, ke kterým ACM a jejich členové vzhlížejí jako k autoritám v době rozvoje informačních technologií [**?**].

2.6.4 ISI Highly Cited highlighted

ISI Highly Cited je databáze často citovaných autorů v článcích posledního desetiletí, které byly vydány institutem ISI (Institute for Scientific Information). Ten v dnešní době spadá pod agenturu Thomson Reuters, na jejíchž webových stránkách nalezneme seznam autorů ISI Highly Cited highlighted z let 2000 až 2008 napříč 21 vědeckými obory [**?**].

KAPITOLA 3

Významnost autorů

3.1 Míry centrality

3.1.1 Degree

Stupeň je počet hran spojených s uzlem. Pro orientovaný graf můžeme uvažovat vstupní (indegree) a výstupní stupeň (outdegree) vrcholu nebo obecný stupeň (degree), tedy součet těchto dvou. Vstupní stupeň se často označuje jako deg^- a výstupní jako deg^+.

$$C_{Din}(u) = deg^-(u) = \sum_{v \in V} \mathbf{A}_{uv} \tag{3.1}$$

$$C_{Dout}(u) = deg^+(u) = \sum_{v \in V} \mathbf{A}_{vu} \tag{3.2}$$

$$C_D(v) = C_{Din} + C_{Dout} \tag{3.3}$$

A je matice sousednosti grafu (adjacency matrix) a prvek této matice $\mathbf{A}_{uv}$ na řádku u a v značí, že existuje hrana z vrcholu u do vrcholu v, je-li hodnota 1, w pokud se jedná o vážený graf a 0, pokud hrana neexistuje.

Pokud uvažujeme pouze vstupní stupeň, vypočtená hodnota určuje významnost uzlu, kdežto výstupní stupeň ukazuje jakousi společenskost či otevřenost uzlu.

Degree centrality je výpočetně velmi jednoduchý způsob, jak změřit významnost prvku v síti. Tato metoda je však příliš jednoduchá, protože do výpočtu hodnoty centrality nezahrnuje uzly, které jsou od daného uzlu vzdálenější než jeden skok. Tento fakt je známý problém a důvod pro zavedení dalších a složitějších metod pro výpočet významnosti.

3.1.2 Eigenvector

Eigenvector centrality, také známá jako Gould's index of accessibility of a Network (Linear Algebra with Applications: Alternate Edition by Gareth Williams), je míra vlivu vrcholu v grafu. Hodnotu vlivu získáme z vlastního vektoru x matice sousednosti grafu:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{3.4}$$

A je matice sousednosti, **x** je vlastní vektor matice **A** a řešením této rovnice o více řešeních. Ke každému řešení náleží vlastní číslo λ. Pro měření významnosti nás však zajímá pouze to řešení, které má pouze nezáporné hodnoty. Podle Perron-Frobeniovy věty pro každou nezápornou primitivní matici existuje právě jedno takové řešení, které zároveň patří k největšímu vlastnímu číslu λ [**?**].

Rovnici můžeme rozepsat z maticového tvaru do jednotlivých složek:

$$x_u = \frac{1}{\lambda}\sum_{v\in G}\mathbf{A}_{uv}x_v \tag{3.5}$$

Kde x_u je prvek vlastního vektoru **x** náležící vrcholu u a $\mathbf{A}_{uv}$ je prvek matice sousednosti **A**, který leží na řádku u a sloupci v.

$$x_{ui+1} = \frac{1}{\lambda}\sum_{v\in G}\mathbf{A}_{uv}x_{vi} \tag{3.6}$$

V tomhle rekurentním tvaru je vidět předpis pro iterační výpočet eigenvector centrality. Algoritmus se nazývá mocninná metoda, která se používá pro řešení problému vlastních čísel v numerické matematice. Výsledkem mocinné metody je dominantní vlastní číslo a odpovídající vlastní vektor. Pro eigenvector centrality nás zajímá právě tohle řešení a žádné jiné.

Z druhé rovnice si navíc povšimneme, že se jedná o přímé rozšíření degree centrality (3.1). Výsledek předchozí iterace použijeme jako vstup do následující a iterujeme tak dlouho, dokud nedosáhneme požadované přesnosti.

PageRank

V roce 1998 vyvinuli Sergey Brin a Larry Page algoritmus PageRank (nesoucí jméno druhého autora) jako součást výzkumu na novém druhu webového vyhledávače (cite něco). PageRank přiřazuje relativní hodnocení webovým stránkám podle hypertextových odkazů z jiných webových stránek, které na ně směřují, a podle jejich PageRankové významnosti. Sama definice je rekurzivní a po nahlédnutí na vzorec zjistíme, že se jedná o rozšířenou variantu algoritmu pro eigenvector centrality.

$$x_{ui+1} = \frac{1-d}{|V|} + d\sum_{v \in V} \mathbf{A}_{uv} \frac{x_{vi}}{deg^+(v)} \tag{3.7}$$

A je opět matice sousednosti, V je množina vrcholů a $deg^+(v)$ je výstupní stupeň vrcholu v. V původní myšlence, kdy PageRank modeluje chování náhodného surfaře, damping factor je pravděpodobnost, že daný surfař přestane opakovaně klikat na odkazy, které najde na webové stránce, na kterou se dostal z předchozího odkazu, a otevře zcela novou stránku, ze které povede podobný sled surfování přes odkazy. Damping factor je často ze zkušenosti nastaven na 85%.

Hodnota PageRanku je z matematického hlediska pravděpodobnost, že surfař, který náhodně kliká na odkazy, se dostane na konkrétní stránku. Součet všech hodnot PageRanku je tedy 1, protože PageRank je rozdělení pravděpodobnosti.

Jedním problémem algoritmu PageRank jsou uzly bez výstupních hran (dangling nodes). Protože musíme v každé iteraci algoritmu zachovat vlastnost rozdělení pravděpodobnosti, že suma všech pravděpodobností je 1, je třeba zajistit, aby se přenášená hodnota mezi iteracemi neztrácela právě v uzlech bez výstupních hran. Problém se nazývá rank sink a nejčastěji se řeší přidáním zdroje PageRanku:

$$x_{ui+1} = \frac{1-d}{|V|} + d\sum_{v \in V} \mathbf{A}_{uv} \frac{x_{vi}}{deg^+(v)} + \frac{1}{|D|}\sum_{w \in D} x_{wi} \tag{3.8}$$

V každé iteraci předem vypočítáme součet hodnot PageRanku, které by se ztratily v uzlech bez výstupních hran. Tahle hodnota je v rámci iterace konstantní a pouze ji rovnoměrně rozdělíme mezi uzly sítě (tedy s váhou $1/|D|$, kde D je množina uzlů bez výstupních hran (dangling nodes).

Přestože je PageRank původně určený pro webovou síť, lze ho použít na jakoukoliv orientovanou váženou i neváženou síť, tedy i na sociální a citační sítě, o kterých je zde řeč. Pro neorientovaný graf je hodnota PageRanku pro jednotlivé uzly velmi blízká stupňům grafu, ale ne totožná (cite icola Perra and Santo Fortunato.; Fortunato (September 2008). "Spectral centrality measures in complex networks")

3.1.3 Míry založené na nejkratších cestách

V sítích dopravní infrastruktury nás zajímá, po které cestě se nejrychleji a nejvýhodněji dostat z bodu A do bodu B. V sociálních a citačních sítích nemůžeme intuitivně hovořit o nějakých cestách mezi uzly, protože ani přesně nevíme jak takovou cestu interpretovat. Nejkratší cesta mezi přáteli v sociální síti může znamenat, přes které přátele se mezi nimi nejpravděpodobněji šíří informace. V sítích spolupráce vědeckých autorů se například setkáme s tzv. Erdősovým číslem, které vyjadřuje nejkratší vzdálenost mezi osobou a matematikem Paulem Erdősem v rámci spolupráce na matematických pracích.

Použijeme-li metody z dopravních sítí pro analýzu sociálních a citačních sítí, které v jádře spočívají v hledání nejkratších cest, setkáme se se dvěma nejznámějšími mírami centrality closeness a betweeness.

Nechť cesta z bodu $u \in V$ do bodu $v \in V$ je střídající se posloupnost vrcholů a hran takových, že spojují předcházející a následující vrchol v této posloupnosti. Délka cesty je pak součet vah hran této cesty nebo pouze počet hran v případě neváženého grafu. Vzdálenost vrcholů $d_G(u, v)$ je délka nejkratší z cest, které spojují vrcholy u a v.

3.1.4 Closeness

Closeness neboli blízkost je definována jako převrácená hodnota míry farness, tedy dalekost. Dalekost je součet všech vzdáleností od uzlu do všech ostatních, tzn. $f(u) = \sum_{v \in V} d_G(u, v)$ a $c(u) = \sum_{v \in V} \frac{1}{d_G(u,v)}$. Podle jiné definice je closeness převrácená hodnota průměrné nejkratší cesty. V podstatě se od předchozí příliš neliší, protože průměrná nejkratší cesta je rovna $\frac{1}{n-1} \sum_{v \in V} d_G(u, v)$ a

closeness podle této definice:

$$c(u) = \frac{n-1}{\sum_{v \in V} d_G(u,v)}$$

Pro obě definice platí, že čím vyšší hodnota $c(u)$, tím je uzel u významnější podle této míry. Zde se budeme soustředit na druhou definici, protože je častou volbou autorů zabývajících se touto problematikou a existuje pro ni aproximační algoritmus, který si zde uvedeme.

Closeness, stejně jako ostatní míry centrality, modelují rozptýlení informace napříč sítí. Výše uvedené klasické definici je vytýkáno, že pro přenos informace uvažuje pouze nejkratší cesty, které nejsou vždy jedinou komunikační cestou v síti. Alternativu navrhli Noh A Rieger (2004), kde namísto nejkratších cest používají náhodné procházky (random walk closeness centrality). Příkladem může být oběh mincí mezi lidmi. Tento jev nemá s nejkratšími cestami mnoho společného, proto je vhodnější ho modelovat náhodnými procházkami. Oproti tomu například poštovní zásilky očividně cestují po nejkratších cestách. Pokud uvažujeme citační sítě, nemáme jasnou představu o významu náhodných procházek nebo nejkratších cest jako v případě mince nebo dopisu. I přesto očekáváme vysokou podobnost této metody s ostatními.

Nevýhodou closeness centrality je nutnost uvažovat souvislý graf, tedy takový, který obsahuje pouze jednu komponentu. Pokud by měl více komponent, pak by vždy existovala cesta s nekonečnou vzdáleností. Hodnota farness by pak byla automaticky nekonečná a closeness, tedy převrácená hodnota, by byla nulová.

Existuje několik upravených definic, které se mají vypořádat s problémem konektivity a druhotně jsou numericky stabilnější. Jedna z nich zaměňuje převrácenou hodnotu součtu vzdáleností za součet převrácených hodnot vzdáleností $c(u) = \sum_{v \in V} \frac{1}{d_G(u,v)}$ (Opsahl) a druhá $c(u) = \sum_{v \in V} 2^{-d_G(u,v)}$ (Dangalchev).

Algoritmus

Closeness pro všechny vrcholy můžeme přesně vypočítat v čase $O(|V||E| + |V|^2 \log |V|)$, kde V a E jsou množiny vrcholů a hran sítě (cite JO77, FT87).

Algoritmus vychází z definice, tedy vyřeší problém všech párů nejkratších cest, čímž rovnou získá hodnoty farness $f(u) = \sum_{v \in V} d_G(u,v)$ a zjištění closeness je poté triviální podle jedné z výše uvedených definic. Výše uvedená

složitost platí pro použití Dijkstrova algoritmu (cite Dijkstra) pro všechny páry cest.

Pro rozsáhlé sítě s miliony uzlů (sociální sítě k dnešnímu datu) je tato metoda příliš náročná. Eppstein a Wang vyvinuli aproximační algoritmus s náročností $O(\frac{\log |V|}{\epsilon}^2(|V| \log |V| + |E|))$ s chybou $\epsilon\delta$ pro převrácenou hodnotu closeness s pravděpodobností alespoň $1 - \frac{1}{n}$, kde $\epsilon > 0$ a δ je diametr sítě (nejdelší z nejkratších cest). Na základě tohoto aproximačního algoritmu byl vytvořen jiný aproximační algoritmus pro nalezení k nejvýznamnějších uzlů hodnocených podle closeness centrality.

Aproximace

Algoritmus TOPRANK (Okamoto, Chen, Li) najde prvních k nejvýznamnějších uzlů s vysokou pravděpodobností a pro každý z nich přesnou hodnotu closeness. Algoritmus pracuje s myšlenkou, že zjistíme přibližné pořadí uzlů tak, že pro jeden strom nejkratších cest nebudeme počítat se všemi koncovými uzly, ale jen s dostatečně velkým vzorkem této množiny. Přesné hodnoty closeness dosáhneme použitím exaktního algoritmu, který použijeme jen na nejvýznamnější uzly získané z prvního aproximovaného kroku. Klíčovou otázkou je kolik nejvýznamnějších uzlů musíme uvažovat, aby se jednalo o dostatečně přesný výsledek. Autoři algoritmu uvádějí tento algoritmus s heuristikou, která najde přibližně místo, ve kterém je vhodné výpočet ukončit a považovat za dostatečně přesný. Sami uvádějí, že tento algoritmus je pouze první krok k návrhu efektivnějšího způsobu jak najít prvních k nejvýznamnějších uzlů.

3.1.5 Betweeness

Betweeness je druhá metoda, která modeluje šíření informace sítí pomocí nejkratších cest. Princip betweeness spočívá v zvýhodnění uzlů s pozicí, přes kterou teče nejvíce informace. Pokud uzel A komunikuje s uzlem C, pak můžeme tvrdit, že uzel B, který leží mezi nimi, bude mít roli prostředníka. Být tímto prostředníkem mezi více takovými uzly intuitivně napovídá, že takový uzel bude centrální. „Čím více lidí na mně závisí k vytvoření spojení s jinými lidmi, tím mám větší moc“ (cite introductino to social network methods). Betweeness měří na kolika nejkratších cestách se uzel nachází. Více se setkáme s definicí, kde do sumy zahrneme poměr cest, na kterých se uzel

nachází, k celkovému počtu cest mezi dvěma uzly (Freeman, 1977; Anthonisse, 1971, Brandes).

$$b(v) = \sum_{s \in V} \sum_{t \in V \setminus s} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$b(v)$ značí hodnotu betweeness centrality pro uzel v, V množinu uzlů, σ_{st} je počet nejkratších cest mezi uzly s a t a $\sigma_{st}(v)$ je počet nejkratších cest, které navíc procházejí uzlem v.

Normalizovaný betweeness je hodnota v intervalu od 0 do 1, kterou získáme tak, že betweeness vydělíme celkovým počtem možných cest - $((|V|-1)(|V|-2))$ pro orientované grafy a $(\frac{(|V|-1)(|V|-2)}{2})$ pro neorientované grafy.

$$b(v) = \frac{1}{(|V|-1)(|V|-2)} \sum_{s \in V} \sum_{t \in V \setminus s} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Vznik betweeness je připisován sociologovi Lintonu Freemanovi (Freeman 77).

Brandesův algoritmus

Ve své práci Ulrik Brandes zmiňuje do té doby nejrychlejší algoritmus pro výpočet betweeness centrality s časovou náročností $\theta(|V|^3)$ a $\theta(|V|^2)$ paměťovými nároky. Tento způsob přistupuje k problému nekratších cest způsobem all-pair shortest paths. Brandesův způsob využívá algoritmů pro nalezení nejkratších cest z jednoho bodu, kde výsledný algoritmus pracuje s paměťovou náročností $O(|V| + |E|)$ a běží v čase $O(|V||E|)$ nebo $O(|V||E| + |V|^2 \log |V|)$ pro nevážený, respektive vážený graf.

Brandes ve své práci o algoritmu uvádí pseudokód pro nevážený graf, který je následně snadné pozměnit pro vážený graf zaměněním obyčejné fronty za prioritní frontu; kompletní důkaz správnosti algoritmu a porovnání standardního algoritmu s tímto (TODO cite Brandes).

Aproximace

I přes rychlejší Brandesův algoritmus je výpočet betweeness centrality příliš náročný výpočet pro sítě reálného světa (např. biologické, dopravní nebo webové sítě) a pokud nám jde více o relativní pořadí uzlů podle hodnoty betweeness než o hodnotu samotnou, lze oželit přesný výpočet přibližným, který příliš nezmění výsledné umístění v žebříčku nejvýznamnějších uzlů.

Bader, Kintali, Madduri, Mihail ukazují aproximační algoritmus pro betweeness a odhadem chyby. Myšlenkou je jednoduchá lineární extrapolace Brandesova algoritmu, pokud do výpočtu zahrneme pouze náhodný vzorek namísto celé množiny vrcholů. Nechť k je velikost vzorku množiny vrcholů, se kterým počítáme, pak extrapolovaná hodnota betweeness je $\frac{|V|S}{k}$, kde S je vypočtená přibližná hodnota.

3.1.6 Radius

3.2 Hledání nejkratších cest

Hledání nejkratších cest v grafu je historicky starý problém, jehož matematický výzkum přišel relativně pozdě v porovnání s jinými problémy kombinatorické optimalizace (nejmenší kostra grafu, přiřazovací a dopravní problém). Pravděpodobně byl výzkum opožděn, protože se jedná o intuitivní a relativně jednoduchý problém, ale jakmile se dostal do středu zájmu, bylo nezávisle na sobě nalezeno několik řešících metod různými lidmi (Shimbel [1955], Ford [1956], Dantzig [1958], Bellman [1958], Moore[1959], Dijkstra [1959]).(cite schrijver alexander ON THE HISTORY OF THE SHORTEST PATH PROBLEM)

Z hlediska metod řešení můžeme uvažovat několik kategorií algoritmů - nalezení všech párů nejkratších cest (all-pairs shortest path problem), nalezení cesty mezi počátečním a koncovým vrcholem (source-target) nebo nalezení stromu nejkratších cest, máme-li zadán počáteční vrchol (single source shortest path problem).

3.2.1 Single source shortest path

Pokud hledáme pouze jednu cestu mezi dvěma vrcholy (source-target), nemusíme počítat celý strom nejkratších cest, ale můžeme zastavit výpočet při dosažení požadovaného vrcholu.

BFS Prohledávání do šířky z jednoho bodu (breadth first search) je algoritmus, který najde nejkratší cesty z jednoho bodu do všech ostatních v případě neváženého grafu v čase $O(E)$.

Bellman-Fordův algoritmus

Dijkstrův algoritmus

3.2.2 All-pair shortest path

Do této kategorie spadají maticové metody, tj. graf je zadán jako matice sousednosti nebo matice sousednost s váhami hran.

Shimbelova metoda [1955] používá upravené maticové násobení k získání $|V|$-té mocniny matice sousednosti. Celková časová náročnost je $O|V|^4$, protože provedeme $|V|$ násobení čtvercové matice o složitosti $O|V|^3$. Shimbelovo upravené násobení nahrazuje sčítání a násobení za minimum a sčítání:

$$x + y \equiv \min(x, y)$$
$$xy \equiv x + y$$

Floyd-Warshallův algoritmus snižuje časovou náročnost na $O|V|^3$ použitím dynamického programování. Graf je opět zadán jako vážená matice sousednosti. Rekurentní vzorec dynamického programování pro tento algoritmus je:

$$d_0(u, v) = \mathbf{A}_{uv}$$
$$d_{k+1}(u, v) = \min(d_k(u, v), d_k(u, k) + d_k(k, v))$$

Jednodušše zkoušíme, zda je kratší cesta mezi vrcholy u a v, kterou již známe, nebo jiná cesta za použití nějakého vrcholu k, který leží mezi nimi. Výpočet provádíme pro všechny páry vrcholů pro všechny vrcholy k ($|V|^2|V|$).

Johnsonův algoritmus nepatří mezi maticové metody, protože využívá metod single source shortest path pro všechny vrcholy. Váhy hran mohou být i záporné, ale v takovém případě je nutné provést transformaci vah pomocí Bellman-Fordova algoritmu, která zachová nejkratší cesty.

3.3 Ostatní používané míry významnosti autorů

3.3.1 H-index

H-index je metoda pro měření produktivity a významu publikované vědecké práce. Metoda byla navržena fyzikem Jorge E. Hirschem pro zjišťování relativní významnosti vědců publikujících v oboru teoretické fyziky. Původní Hirschova definice zní:

> Vědec má index h pokud h z jeho N_p publikací má každá alespoň h citací a žádná ze zbylých $(N_p - h)$ publikací nemá více než h citací.

Tato metoda je specifická pro citační sítě a nelze ji aplikovat na sociální nebo obecnou komplexní síť.

Výpočet probíhá tak, že seřadíme autorovy publikace sestupně podle počtu citací a poté od začátku tohoto seřazeného seznamu hledáme tu publikaci, jejíž počet citací je nižší než pořadové číslo v tomto seznamu. Toto číslo je pak h-index autora.

3.3.2 Impact factor

3.4 Porovnání výsledků

Pro zjištění, jestli jsou výsledky implementovaných metod shodné s uvedenými oceněními, použijeme metodu součtu pořadí oceněných autorů. Tzn. pro jedno ocenění sečteme pořadí všech autorů, kteří byli oceněni touhle cenou. Tuhle jednoduchou míru můžeme porovnat pouze mezi jednotlivými metodami pro jedno ocenění, ale ne mezi různými oceněními pro jednu metodu. Jednodušše protože například Turingova cena je ve výsledcích udělena pouze několika autorům, z čehož plyne malý součet pořadí, kdežto velké množství autorů

na prvních pozicích je členy ACM Fellows, tím pádem velký součet pořadí oceněných.

Dále nás zajímá, zda-li jsou metody mezi sebou podobné či nikoliv. Pro porovnání metod mezi sebou je použit Spearmanův koeficient pořadové korelace.

3.4.1 Spearmanův koeficient pořadové korelace

Spearmanův koeficient je klasický Pearsonův koeficient korelace, který je aplikovaný na proměnné s pořadím (cite Myers, Jerome L.; Well, Arnold D. (2003), Research Design and Statistical Analysis (2nd ed.), Lawrence Erlbaum, pp. 508).

Koeficient korelace obecně dosahuje hodnot od -1 do 1, přičemž hodnota 1 znamená naprostou lineární závislost mezi porovnávanými proměnnými tak, že s rostoucí první proměnnou roste i druhá proměnná. Hodnota -1 znamená rovněž naprostou lineární závislost, ale při rostoucí jedné proměnné druhá proměnná klesá. Hodnota 0 znamená kompletní náhodnost či nezávislost mezi měřenými proměnnými.

KAPITOLA 4

Výsledky

4.1 Porovnání nejvýznamnějších autorů

4.2 Porovnání implementovaných metod

KAPITOLA 5

Diskuse

5.1 Podobnost výsledků jednotlivých metod

5.2 Shoda výsledků s oceněními

5.3 Vliv vah na přesnost výsledků

5.4 Vstupní a výstupní hrany

KAPITOLA 6

Závěr

PŘÍLOHA A

Žebříčky významných autorů

A.1 DBLP

	Autor	outdegree	Turing	Codd	Fellows	ISI
1	GERHARD WEIKUM	872.000			•	
2	HECTOR GARCIA-MOLINA	856.000		•		
3	RAKESH AGRAWAL	761.000		•		
4	MICHAEL J. CAREY	758.000		•		
5	DAVID J. DEWITT	758.000		•	•	
6	H. V. JAGADISH	717.000			•	
7	MICHAEL STONEBRAKER	677.000		•	•	
8	RAGHU RAMAKRISHNAN	652.000				
9	YANNIS E. IOANNIDIS	649.000				
10	ABRAHAM SILBERSCHATZ	636.000				
11	ELISA BERTINO	635.000			•	
12	SHAMKANT B. NAVATHE	629.000				
13	PHILIP S. YU	622.000				
14	STEFANO CERI	611.000				
15	CHRISTOS FALOUTSOS	607.000				
16	MATTHIAS JARKE	586.000				
17	GULTEKIN OUMLZSOYOGLU	582.000				
18	SERGE ABITEBOUL	575.000		•	•	•
19	NICK ROUSSOPOULOS	568.000				
20	MIRON LIVNY	559.000				
21	STANLEY Y. W. SU	558.000				
22	HANS-JOUMLRG SCHEK	557.000			•	
23	PATRICK VALDURIEZ	547.000				
24	GOETZ GRAEFE	546.000				
25	CLEMENT T. YU	542.000				
26	RICHARD HULL	537.000			•	
27	MICHAEL J. FRANKLIN	526.000				
28	RICHARD T. SNODGRASS	513.000			•	
29	JENNIFER WIDOM	510.000		•	•	
30	DENNIS SHASHA	508.000				

	Autor	**indegree**	**Turing**	**Codd**	**Fellows**	**ISI**
1	MICHAEL STONEBRAKER	1909.000		•	•	
2	DAVID J. DEWITT	1484.000		•	•	
3	JIM GRAY	1400.000		•		
4	RAYMOND A. LORIE	1276.000				
5	JEFFREY D. ULLMAN	1180.000		•		
6	WON KIM	1146.000				
7	PHILIP A. BERNSTEIN	1145.000		•		
8	E. F. CODD	1110.000		•		
9	MICHAEL J. CAREY	1110.000		•		
10	UMESHWAR DAYAL	1076.000		•		
11	HECTOR GARCIA-MOLINA	1020.000		•		
12	DAVID MAIER	1017.000		•		
13	DONALD D. CHAMBERLIN	966.000		•	•	
14	RAKESH AGRAWAL	907.000		•		
15	PETER P. CHEN	906.000				
16	SERGE ABITEBOUL	848.000		•	•	•
17	KAPALI P. ESWARAN	847.000				
18	MORTON M. ASTRAHAN	846.000				
19	FRANCCEDILOIS BANCILHON	840.000				
20	NATHAN GOODMAN	819.000				
21	BRUCE G. LINDSAY	806.000				
22	HAMID PIRAHESH	803.000			•	
23	IRVING L. TRAIGER	785.000			•	
24	EUGENE WONG	762.000				
25	JEFFREY F. NAUGHTON	729.000				
26	JENNIFER WIDOM	727.000		•	•	
27	RAGHU RAMAKRISHNAN	724.000				
28	CATRIEL BEERI	722.000			•	•
29	NICK ROUSSOPOULOS	702.000				
30	SHAMKANT B. NAVATHE	694.000				

	Autor	**pagerank**	**Turing**	**Codd**	**Fellows**	**ISI**
1	E. F. CODD	179.324		•		
2	MICHAEL STONEBRAKER	137.371		•	•	
3	JIM GRAY	115.364		•		
4	DONALD D. CHAMBERLIN	114.010		•	•	
5	RAYMOND A. LORIE	107.204				
6	PHILIP A. BERNSTEIN	99.575		•		
7	MORTON M. ASTRAHAN	87.673				
8	KAPALI P. ESWARAN	87.167				
9	PETER P. CHEN	84.098				
10	IRVING L. TRAIGER	79.313			•	
11	JOHN MILES SMITH	78.833				
12	JEFFREY D. ULLMAN	74.323		•		
13	EUGENE WONG	68.319				
14	DAVID J. DEWITT	67.701		•	•	
15	MIKE W. BLASGEN	62.185				
16	GIANFRANCO R. PUTZOLU	61.585				
17	BRADFORD W. WADE	60.731				
18	RUDOLF BAYER	60.706		•		
19	JAMES W. MEHL	58.499				
20	PATRICIA P. GRIFFITHS	58.215				
21	WON KIM	57.946				
22	W. FRANK KING III	57.169				
23	NATHAN GOODMAN	56.791				
24	PAUL R. MCJONES	55.967			•	
25	RONALD FAGIN	54.766		•	•	•
26	RAYMOND F. BOYCE	54.475				
27	UMESHWAR DAYAL	54.099		•		
28	DIANE C. P. SMITH	53.677				
29	VERA WATSON	53.085				
30	MICHAEL HAMMER	52.687				

	Autor	**inCloseness**	**Turing**	**Codd**	**Fellows**	**ISI**
1	MICHAEL STONEBRAKER	0.593		•	•	
2	JIM GRAY	0.560		•		
3	DAVID J. DEWITT	0.556		•	•	
4	RAYMOND A. LORIE	0.556				
5	JEFFREY D. ULLMAN	0.546		•		
6	PHILIP A. BERNSTEIN	0.546		•		
7	E. F. CODD	0.543		•		
8	DONALD D. CHAMBERLIN	0.539		•	•	
9	WON KIM	0.537				
10	UMESHWAR DAYAL	0.535		•		
11	MICHAEL J. CAREY	0.532		•		
12	MORTON M. ASTRAHAN	0.531				
13	DAVID MAIER	0.529		•		
14	KAPALI P. ESWARAN	0.529				
15	NATHAN GOODMAN	0.527				
16	EUGENE WONG	0.526				
17	IRVING L. TRAIGER	0.525			•	
18	HECTOR GARCIA-MOLINA	0.523		•		
19	FRANCCEDILOIS BANCILHON	0.520				
20	BRUCE G. LINDSAY	0.519				
21	PETER P. CHEN	0.518				
22	RAKESH AGRAWAL	0.518		•		
23	RONALD FAGIN	0.517		•	•	•
24	CATRIEL BEERI	0.517			•	•
25	THOMAS G. PRICE	0.514				
26	PATRICIA G. SELINGER	0.514		•	•	
27	JOHN MILES SMITH	0.513				
28	MIKE W. BLASGEN	0.512				
29	RANDY H. KATZ	0.512				
30	GIO WIEDERHOLD	0.512				

	Autor	outCloseness	Turing	Codd	Fellows	ISI
1	H. V. JAGADISH	0.475			•	
2	RAKESH AGRAWAL	0.473		•		
3	GERHARD WEIKUM	0.471			•	
4	HECTOR GARCIA-MOLINA	0.468		•		
5	GULTEKIN OUMLZSOYOGLU	0.467				
6	YANNIS E. IOANNIDIS	0.467				
7	STEFANO CERI	0.466				
8	SHAMKANT B. NAVATHE	0.466				
9	MICHAEL J. CAREY	0.465		•		
10	ELISA BERTINO	0.464			•	
11	RAGHU RAMAKRISHNAN	0.463				
12	RICHARD T. SNODGRASS	0.463			•	
13	DAVID J. DEWITT	0.462		•	•	
14	SERGE ABITEBOUL	0.462		•	•	•
15	CLEMENT T. YU	0.460				
16	GOETZ GRAEFE	0.460				
17	HANS-JOUMLRG SCHEK	0.459			•	
18	ABRAHAM SILBERSCHATZ	0.459				
19	JENNIFER WIDOM	0.459		•	•	
20	PATRICK VALDURIEZ	0.458				
21	NICK ROUSSOPOULOS	0.457				
22	RICHARD HULL	0.457			•	
23	UMESHWAR DAYAL	0.457		•		
24	MICHAEL STONEBRAKER	0.454		•	•	
25	DENNIS SHASHA	0.454				
26	MATTHIAS JARKE	0.453				
27	MIRON LIVNY	0.451				
28	HAMID PIRAHESH	0.451			•	
29	CHRISTIAN S. JENSEN	0.451				
30	ALBERTO O. MENDELZON	0.450				

A.2 CiteSeer