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

Bakalářská práce

Měření významnosti autorů v citační síti

Plzeň, 2013

Tomáš Maršálek

Abstract

Obsah

KAPITOLA 1

Úvod

KAPITOLA 2

Sociální a citační sítě

2.1 Sociální sítě

2.2 Analýza sociálních sítí

2.3 Citační sítě

Citační sítě jsou speciálním případem sociálních sítí, kde jako uzly uvažujeme knihy, články nebo jiné publikace; nebo uzly mohou být samotní autoři těchto publikací. Spojení mezi uzly je určeno citacemi na jednotlivé publikace.

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.

2.4 Významnost autorů

2.5 Míry centrality

2.5.1 Degree

2.5.2 Eigenvector

2.5.3 Closeness

2.5.4 Betweeness

2.6 Ostatní míry významnosti autorů

2.6.1 H-index

2.7 Metody porovnaní

2.7.1 Spearmanův koeficient pořadové korelace

2.7.2 Pearsonův korelační koeficient

2.8 Ocenění významných autorů

2.8.1 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].

2.8.2 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 [**?**].

2.8.3 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 [**?**].

2.8.4 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

Implementace

3.1 Načtení vstupních dat

3.2 Vytvoření citačních sítí

3.3 Analýza struktury sítě

3.4 Knihovna pro SNA

3.4.1 Reprezentace citační sítě

3.4.2 Degree

3.4.3 Eigenvector

3.4.4 Closeness

3.4.5 Betweeness

3.4.6 H-index

KAPITOLA 4

Výsledky

4.1 Citační databáze

4.1.1 DBLP

DBLP [?] je webová bibliografická databáze v oboru informatiky, která k listopadu 2012 obsahovala více než 2,1 milionu publikací. Pro tuto práci používáme verzi z roku 2004.

4.1.2 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.

4.2 Struktura sítě

4.2.1 DBLP

4.2.2 CiteSeer

4.3 Žebříčky významných autorů

4.3.1 DBLP

H-index

	Autor	H-index	Turing	Codd	Fellows	ISI
1	MICHAEL STONEBRAKER	28.000		•	•	
2	DAVID J. DEWITT	24.000		•	•	
3	JEFFREY D. ULLMAN	24.000		•	•	•
4	PHILIP A. BERNSTEIN	22.000		•	•	•
5	RAKESH AGRAWAL	21.000		•	•	
6	WON KIM	21.000		•	•	
7	YEHOSHUA SAGIV	20.000				
8	CATRIEL BEERI	20.000			•	•
9	MICHAEL J. CAREY	20.000		•	•	
10	SERGE ABITEBOUL	19.000		•	•	•
11	HECTOR GARCIA-MOLINA	19.000		•	•	•
12	UMESHWAR DAYAL	19.000		•	•	
13	CHRISTOS FALOUTSOS	19.000			•	•
14	NATHAN GOODMAN	18.000		•		
15	JIM GRAY	18.000			•	
16	JEFFREY F. NAUGHTON	18.000			•	
17	RAGHU RAMAKRISHNAN	18.000				
18	RONALD FAGIN	18.000		•	•	•
19	JENNIFER WIDOM	18.000		•	•	
20	DAVID MAIER	17.000		•	•	•
21	BRUCE G. LINDSAY	17.000			•	
22	SHAMKANT B. NAVATHE	16.000				
23	C. MOHAN	16.000		•	•	
24	HAMID PIRAHESH	16.000			•	
25	H. V. JAGADISH	16.000			•	

Degree

PageRank

Hodnoty PageRanku dosahují hodnot mezi 0 a 1. Pro účely přehlednosti byly v této tabulce normalizovány na interval 0 až $|V|$, tedy počet uzlů sítě.

	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		•	•	•

Closeness

Betweeness

H-index

4.4 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