Project Presentation

Using PageRank to Rank Conferences, Articles, and Authors

Yavar Naddaf

Motivation: Why Rank Conferences?

- We need methods to compare the impact of research done by scientists
  - Examples:
    - Hiring or promoting researchers
    - Assigning funding to different research groups
- Publications are often the only direct output of academic research
- In Computer Science, most of important research is originally published in conferences

Impact Factor

- One of the most commonly used measures to rank journals
- Average number of citations that each article published in a journal receives in a two years period.
- Bollen et al. Argue that it is a measure of Popularity and not Prestige of journals*

* J. Bollen, M. Rodriguez, H. Van de Sompel
Journal Status. Scientometrics, Volume 69, n3, pp 669-687, 2006

PageRank

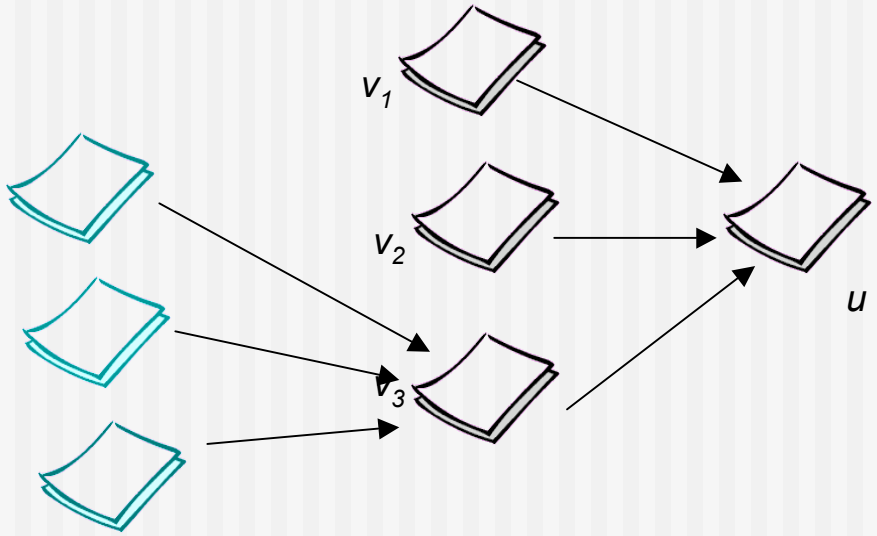
- PageRank* calculates a measure of importance for web pages based on the link structure of the web

- The importance of a page is not only based on the number of other pages that link to it, but also their importance

* Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project

PageRank

- Rank Is Recursive

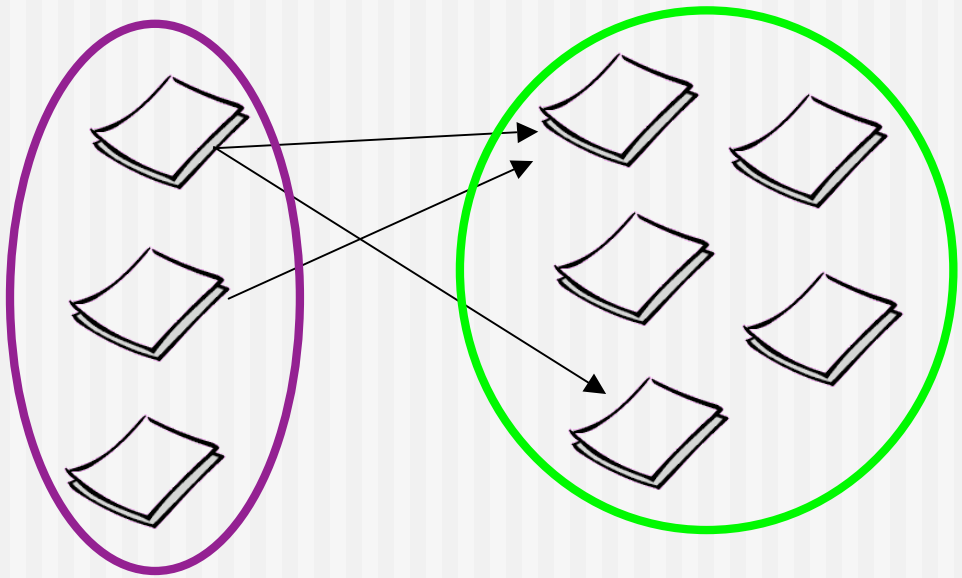


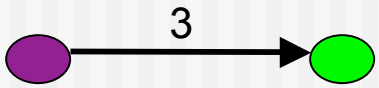
$$Rank(u) = c \sum_{v \in B_u} \frac{Rank(v)}{N_v}$$

* Needs some more details to converge

Journal Citation Network

- All papers published in a journal are presented as one node
- The weight of the edge between j_1 and j_2 is the number of papers in j_1 that cite a paper in j_2.





Journal PageRank

- Similar to PageRank, but applied to Journal Citation Network

- Unlike regular PageRank, the rank of a node is not distributed uniformly among the nodes it links to.

$$Rank(v_j) = c \sum_{v_k} Rank(v_k) w(v_k, v_j)$$

* Needs some more details to converge

Conference PageRank

- Triviality easy to apply the same algorithm on the conference citation network
- The only non-trivial part is generating the citation network (more about this later).

Conference PageRank

- In Matrix Notation
  - For *n* conferences, we build an *n x n* matrix *A, where*:

$$A[i,j] = \frac{C(i,j)}{\sum_k C(i,k)}$$

  - *C(i, j)* is the the number of papers published in conference *i* that cite a paper published in conference *j*.

Conference PageRank

- Start with an arbitrary *1 x n* vector *r*
- *Update r:*

$$r = (1 - d)rA + dq$$

- *rA* is the same recursive rank propagation
- *dq* is a random jump with probability *d*
  - It guarantees that this recursive formula will converge
  - $q_{1 \times n} = [1/n\ 1/n\ \ldots\ 1/n]$

Conference PageRank

- *Repeat updating:*

$$r = (1 - d)rA + dq$$

until the change in *r* is smaller than some predetermined threshold.

- *We used a threshold of* $|r_{t+1} - r_t| < 10^{-5}$
- *d = 0.15*
- The formula always converged within *20* iterations

Conference PageRank

- *Dealing with self-citations*
  - Idea: citations from outside a conference are more important than citations from papers within the same conference
  - We can multiply the main diagonal entries of A by a weight $w \leq 1$
  - We solved two versions of A
    - $w = 1.0$: self-citations fully counted
    - $w = 0.5$: self-citations are discounted by 50%

Article PageRank

- The rank of conferences is used to compare articles (or authors)
- This is controversial:
  - Articles published in a publication have a high variance in the number of cites received
  - A paper published in a high ranking conference is not necessarily of high rank

Article PageRank

- We would like to rank articles directly
- We can apply the exact same algorithm on the Article Citation Network
  - For *n* articles, we build an *n x n* matrix *A, where*:

$$A[i,j] = \begin{cases} \frac{1}{N_i} & \text{if article } i \text{ cites article } j \\ 0 & \text{otherwise} \end{cases}$$

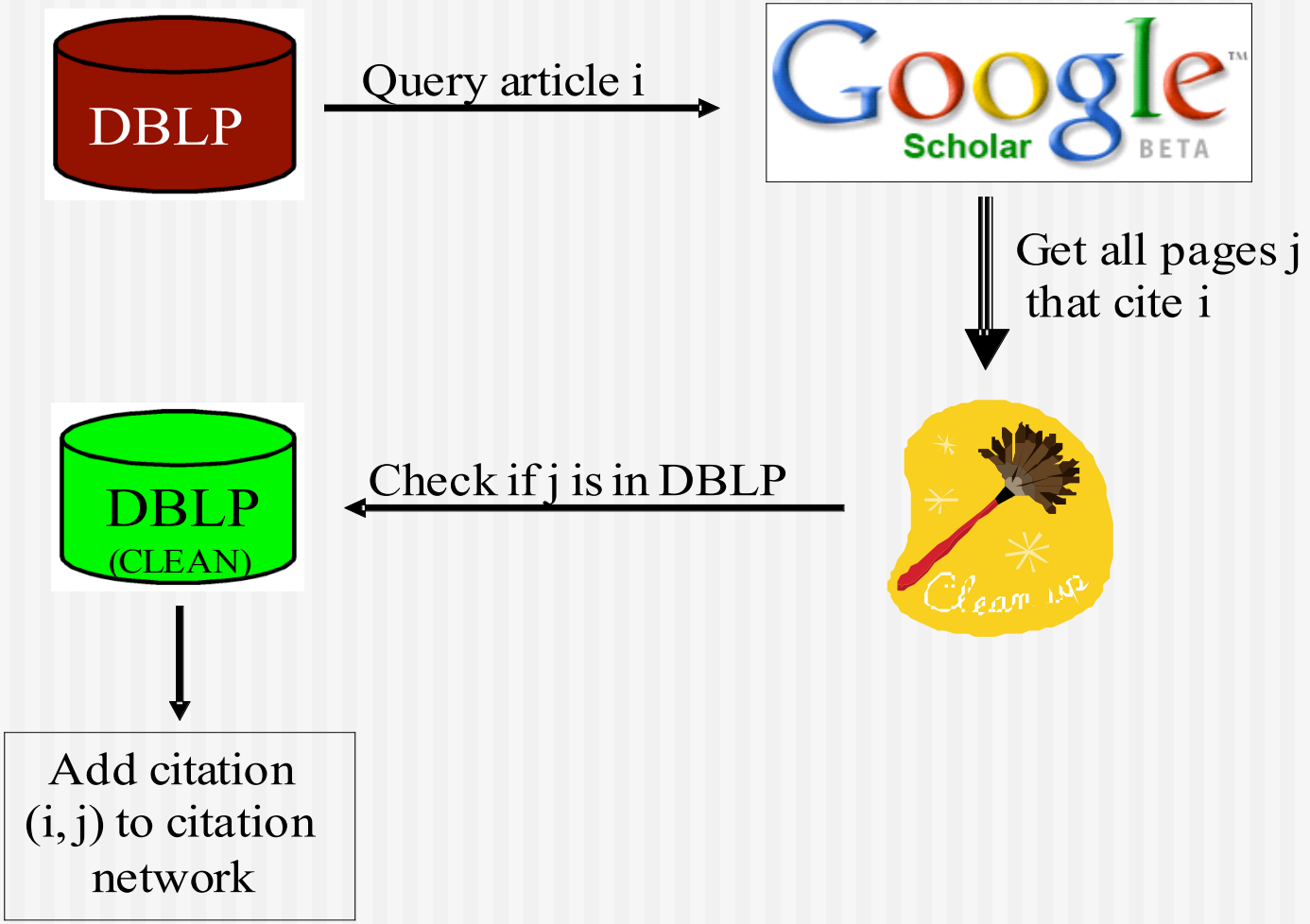
Author PageRank

- We can also rank researchers directly
  - For *n* authors, build an *n x n* matrix *A, where*:

$$A[i,j] = \frac{C(i,j)}{\sum_k C(i,k)}$$

    - *C(i, j)* is the the number of papers written by author *i* that cite a paper written by author *j*.

Extracting the Citation Network



Extracting the Citation Network

- Extracted the citation network for *53,151* articles
  - Published between 2004-2006
  - In *1,682* Database, Machine Learning, AI related conferences
- Generated the citation matrices for conferences, articles, and authors

Sample Rankings

- Database conferences:

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
SIGMOD	12.0×10^{-3}	11.9×10^{-3}	6.51×10^{-6}	1.74
VLDB	10.5×10^{-3}	10.5×10^{-3}	5.06×10^{-6}	1.52
ICDE	8.13×10^{-3}	8.22×10^{-3}	5.23×10^{-6}	1.25
DEXA	0.14×10^{-3}	0.14×10^{-3}	3.08×10^{-6}	0.27
IDEAS	0.26×10^{-3}	0.27×10^{-3}	3.12×10^{-6}	0.27

Sample Rankings

- Machine Learning conferences:

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
KDD	3.37×10^{-3}	3.26×10^{-3}	4.72×10^{-6}	1.68
ICDM	0.66×10^{-3}	0.69×10^{-3}	3.32×10^{-6}	0.35
SDM	1.21×10^{-3}	1.26×10^{-3}	4.04×10^{-6}	0.62
ADMA	0.09×10^{-3}	0.09×10^{-3}	3.00×10^{-6}	N/A

Sample Rankings

- AI conferences:

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
IJCAI	2.04×10^{-3}	2.10×10^{-3}	3.68×10^{-6}	1.10
AAAI	2.66×10^{-3}	2.63×10^{-3}	3.44×10^{-6}	1.49
ICML	2.89×10^{-3}	2.55×10^{-3}	4.30×10^{-6}	2.12
Canadian Conference on AI	0.11×10^{-3}	0.10×10^{-3}	3.03×10^{-6}	0.26

Sample Rankings

- Some members of the GAMES Group

Author Name	Author PageRanak
Jonathan Schaeffer	50×10^{-6}
Russell Greiner	31×10^{-6}
Duane Szafron	30×10^{-6}
Michael Buro	11×10^{-6}
Vadim Bulitko	8×10^{-6}
Michael Bowling	8×10^{-6}
Robert Holte	4×10^{-6}

Limitations and Future Work

- Citation Network is incomplete
  - Add citations from CiteSeer and other sources
- Is PageRank preferable to alternative rankings (like IF)?
  - More analysis is required