Bringing Order to Science

Dimitri Bouche & Cyril Verluise

April 26th, 2018

Abstract

Keywords:

> The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous (...) However unlike flat document collections the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages such as link structure (...) we take advantage of the link structure of the Web to produce a global "importance" ranking of every web page
>
> *L. Page, S. Brin, R. Motwani and T. Winograd (1999)*

Contents

1 Related literature

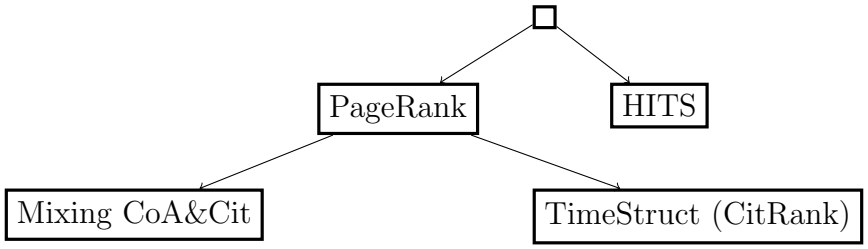


Figure 1: Wrap-up

1.1 Foundations

Brin and Page (1998) - PageRank: In order to measure the relative importance of web pages, we propose PageRank, a method for computing a ranking for every web page based on the graph of the web.

Defnition 1: Let E(u) be some vector over the Web pages that corresponds to a source of rank. Then, the PageRank of a set of Web pages is an assignment, R0, to the Web pages which satisfies:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

such that c is maximized and $||R0||_1 = 1$.

Computing PageRank:
$\texttt{R0} \leftarrow \texttt{S; loop : ; R}_{i+1} \leftarrow \texttt{AR}_i \texttt{ ; d} \leftarrow \texttt{||R}_i\texttt{||}_1 \texttt{ - ||R}_{i+1}\texttt{||}_1 \texttt{ ; R}_{i+1} \leftarrow \texttt{R}_{i+1} \texttt{ + dE ; } \delta \leftarrow \texttt{||R}_{i+1} \texttt{ - R}_i\texttt{|| ; while } \delta > \epsilon$

Where $A_{u,v} = \frac{1}{N_u}$ if there is an edge from u to v, 0 otherwise. Note that the d factor increases the rate of convergence.

Most experiments are performed with a uniform E over all pages. However, another extreme is to have E consists of a single web page. See section "Personalized PageRank".

Kleinberg (1999) - HITS: The author develop a set of algorithmic tools for extracting information from the link structures of such environments. The central issue he addresses within is the distillation of broad search topics, through the

discovery of "authoritative" information sources on such topics. He proposes and tests an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of "hub pages" that join them together in the link structure.
Set-up: Suppose we are given a broad-topic query, specified by a query string σ. We wish to determine authoritative pages by an analysis of the link structure; but first we must determine the subgraph of the www on which our algorithm will operate. Our goal here is to focus the computational effort on relevant pages. Main algorithmic tools are :

- *Focused Subgraph*: we could restrict the analysis to the set Q_σ of all pages containing the query string; but this has two significant drawbacks. First, this set may contain well over a million pages, and hence entail a considerable computational cost; and second, we have already noted that some or most of the best authorities may not belong to this set.

    1. We first collect the t highest ranked pages for the query s from a text-based search engine. The *root set* - R_σ ($R_\sigma \subset Q_\sigma$) is **relatively small** and **rich in relevant pages** but might not **contain most of the strongest authorities** (Holy trinity)
    2. We can use the root set R_σ, however, to produce a set of pages S_σ that will satisfy the conditions we 3 are seeking. A strong authority for the query topic—although it may well not be in the set R_σ, it is quite likely to be pointed to by at least one page in R_σ.

- *Hubs* and *authorities*

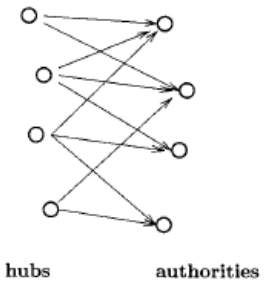


Figure 2: A densely linked set of hubs and authorities

Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

THEOREM 3.2. (SUBJECT TO ASSUMPTION (†)).Let A be the adjacency matrix of a given network, x^* the vector of authorithy weights and y^* the vector of hub weights. x^* is the principal eigenvector of A^TA, and y^* is the principal eigenvector of AA^T.

See p 609 and 613 for pseudo code of relevant algorithms and p 620 for a comparison with PageRank.

Extensions of the HITS algorithm can be found in Lempel and Moran (2000).

1.2 PageRank extensions for citations and co-authors

1.2.1 Mixing citations and co-authors

Yan and Ding (2011) : This article provides an alternative perspective for measuring author impact by applying PageRank algorithm to a coauthorship network. A weighted PageRank algorithm considering citation and coauthorship network topology is proposed. PR_W provides an integrated algorithm to combine citation and the topology of the network in a simple and efficient way. In one

Baseline Pagerank is given by :

$$PR(p) = \frac{1-d}{N} + d\sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)}$$

where N is the total number of pages on the web, p_i is the page that links to p, d is damping factor, and $C(p_i)$ is the number of outlinks of p_i.

Authors' idea is the following: "influential authors should have a better chance to be randomly surfed". Thus, they incorporate citation counts with topology of network:

$$PR_W(p) = (1-d)\frac{CC}{\sum j = 1^N CC(p_j)} + d\sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)}$$

where p is now an *author*, $CC(p)$ is the number of citations pointing to author p, $\sum j = 1^N CC(p_j)$ is the citation counts of all nodes in the network.

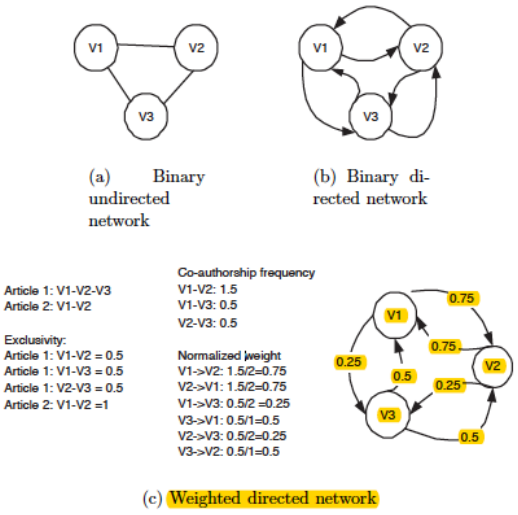
See p6 for implementation of the algorithm.

Fiala et al. (2008): this article proposes a new verison of PageRank which incorporates both citation and co-authorship graph property. They change $\frac{1}{C(P)}$ to $\frac{\sigma}{\sum\sigma}$ wher σ is a value between author i and j and $\sum\sigma$ is the sum of σ over all authors. Embedded in σ is another fraction which measures the numbers of citations from author i and j and all citations from author i to the rest of the authors.

See pp 136-137 for definitions.

Ndr: Same spririt as Yan and Ding (2011) but modifies the second of the PR computation. Certainly computationally *very* intensive

Liu et al. (2005) - AuthorRank: This article introduces a weighted directional network model to represent the co-authorship network. Using this topology, the authors define AuthorRank as an indicator of the impact of an individual author in the network. The results show clear advantages of PageRank and AuthorRank over degree, closeness and betweenness centrality metrics. The article also provide a plethora of descriptive statistics and performs validation methods (spearman correlation, etc).



(a) Binary undirected network

(b) Binary directed network

(c) Weighted directed network

Figure 3: Co-authorship network

The AuthorRank can be calculated with the same iterative algoirthm used by PageRank. One may think of AuthorRank as a generalization of PageRank by substituting $w_{j,i}$ with $\frac{1}{C(j)}$ in PageRank in which $C(j)$ is the number of links going out of page j.

Formally, we have :

$$AR(i) = (1-d) + d\sum_{j=0}^{n} AR(j) * w_{j,i}$$

where $AR(j)$ is the AR value of the backlinking node and $w_{j,i}$ corresponds to the edge weight between node j and i.

Notations: n authors denoted as $V = v_1, \ldots, v_n$. m articles denoted as $A = a_1, \ldots, a_m$ and $f(a-k)$ the number of authors of article a_k. We define:

1. a measure of *exclusivity* $g_{i,j,k} = \frac{1}{f(a_k)-1}$ the degree to which author v_i and v_j have an exclusive co-authorship relation for a particular article
2. a measure of *co-authorship frequency* $c_{ij} = \sum_{k=1}^{m} g_{i,j,k}$ which is a way to give more weight to authors who co-publish more papers together and do so exclusively
3. a normalized weight $w_{i,j} = \frac{c_{i,j}}{\sum_{k=1}^{n}}$ where the normalization ensure that the weights of an author sum to 1.

See pp 7-8, 11 for technical details.

1.2.2 Adressing time structure of citations graph

Walker et al. (2007); Maslov and Redner (2008) - CiteRank: This article modifies the PageRank algorithm to account for the "time censorship" of citations graphs. The authors' solution is to initially distribute random surfers exponentially with age, in favour of more recent publications. The algorithm is called *CiteRank*.

Formally, they define:

1. $W_{i,j} = \frac{1}{k_j^{out}}$ if j cites i and 0 otherwise where k_j^{out} is the out degree of paper j
2. $\rho_i = e^{-\frac{age_i}{\tau_{dir}}}$ and the probability that a researcher finds a given paper by initial selection is given by ρ. Intuitively, $1/\alpha$ refers to the average depth of a citation chain and τ_{dir} refers to the average age of the initial article of the chain.

The CiteRank is defined as the proba of encountering a paper via a paths of any length, ie, the CiteRank traffic is given by :

$$T = I.\rho + (1-\alpha)W.\rho + (1-\alpha)^2 W^2.\rho + \dots$$

Practically, they calculate the CiteRank traffic on all papers in the dataset by taking successive terms in the above expansion to sufficient convergence ($< E10$ of the average value).
The authors find alpha $\alpha \approx 0.5$ and $\tau_{dir} \in [1;3]$ optimal to match effective traffic of their databases.

1.3 Other papers and results of interest

London et al. (2015) - Local PageRank estimation : This article defines a modified PageRank algorithm and the PR-score to measure the influence of a single article by using its local co-citation network.

1. Subgraph building: Starting form certain target nodes (articles), for which we are interested in measuring their scientific impact, and expanding backward by following reversely the nodes having out-going links to the target nodes. The procedure stops after a fixed number of levels. This can be done by an iterative deepening depth-first search. In this work, the graphs contain all nodes, from which the target nodes can be reached in in at most three steps and we consider the induced subgraph of that nodes. Nb: radius is set to 3

2. Estimating the PR of the boundary: We use a heuristic to estimate the individual PR: in each iteration turn, we add an extra term to the PR value of each boundary node that equals to the fraction of its in-coming edges to all edges in the subgraph.

3. Calculating the PR and RP: On one hand, we run the PageRank algorithm on the subgraph, in each step we use the estimated PR value of the boundary nodes adding the λ/N damping factor to each node. On the other hand, we also calculate the reaching probability, RP, of the target node(s) in the subgraph.

See pp 4-5 for technical details

Chen et al. (2007) - Identifying scientific "gems" using PageRank : This article finds out scientific *gems* using the ratio between the Pagerank value and the number of citations. While the Google number and the number of citations for each publication are positively correlated, outliers from this linear relation identify some exceptional papers or "gems" that are universally familiar to physicists.

Fyi:

- Ranks produced by Pagerank and PageRank derivatives should be robust to the choice of d (damping factor)
- Brin and Page (1998) choose d=O.15 but alternatives focused on citations and co-authorship network argue that d should be closer to 0.5. Heuristically, research depth of scientists is said to be closer to 2 than 6.
- To compare results of different algorithms, usual practice is to use the *Spearman correlation coefficient* and/or *QQ-plots*.
- Rank-aggregation ("bagging") is usually made via Kemeny voting algorithm.

References

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15.

Fiala, D., Rousselot, F., and Ježek, K. (2008). Pagerank for bibliographic networks. *Scientometrics*, 76(1):135–158.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Lempel, R. and Moran, S. (2000). The stochastic approach for link-structure analysis (salsa) and the tkc effect1. *Computer Networks*, 33(1-6):387–401.

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing & management*, 41(6):1462–1480.

London, A., Németh, T., Pluhár, A., and Csendes, T. (2015). A local pagerank algorithm for evaluating the importance of scientific articles. In *Annales Mathematicae et Informaticae*, volume 44, pages 131–141. szte.

Maslov, S. and Redner, S. (2008). Promise and pitfalls of extending google's pagerank algorithm to citation networks. *Journal of Neuroscience*, 28(44):11103–11105.

Walker, D., Xie, H., Yan, K.-K., and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010.

Yan, E. and Ding, Y. (2011). Discovering author impact: A pagerank perspective. *Information processing & management*, 47(1):125–134.

Appendix