Authoritative Sources in a Hyperlinked Environment

JON M. KLEINBERG (PAPER)
MARTIN CETKOVKÝ (PRESENTATION)

http://martin.alikuvkoutek.cz

Table of contents

- Introduction
- Graph model
- Hubs and authorities
- Iterative algorithm
- Similar-Page queries
- Related work
- Evaluation & Summary
- Discussion
Introduction

- **Searching on the www**
  - Process of discovering pages relevant to a given query
- **Quality of a search method**
  - Requires human evaluation (*relevance* is subjective)
  - Lack of objective functions that are concretely defined and correspond to human notions of quality.

- The Paper is dated 1999

Problem – Query types

- **Specific queries**
  - “Does Microsoft support Linux?”
- **Broad topic queries**
  - “Find information about the Java programming language.”
- **Similar-page queries**
  - “Find pages ‘similar’ to ‘java.sun.com’.”
Text based search – Limitation

Hyperlink

- Link from $p$ to $q$
  - $p$’s author has in some measure conferred authority on $q$
- Pitfalls
  - Navigational links
    - “(Finally) Go Home”
  - Advertisements
    - “Visit our sponsor”
Graph theory

- Directed graph
  - $G = (V, E)$
- Out-degree
  - $\text{deg}_{\text{out}}(p) = |\{(p, q); (p, q) \in E\}|$
- In-degree
  - $\text{deg}_{\text{in}}(p) = |\{(q, p); (q, p) \in E\}|$
- Induced subgraph
  - $G[W] = (W, F), W \subset V, E \cap (W \times W); G[W] \subset G$

Directed graph – Example
Graph model

- Hyperlinked pages
  - Directed graph \( G = (V, E) \)
  - Nodes \( V \sim \) pages
  - Edges \( E \sim \) links

- Links from the page \( p \)
  - \( \text{deg}_{\text{out}}(p) \)

- Links to the page \( p \)
  - \( \text{deg}_{\text{in}}(p) \)

Hubs and Authorities

- Authorities
  - Most important pages related to the query
  - Linked by many hubs

- Hubs
  - Pages that link to many related authorities
Focused graph

- Focus on relevant pages only
- Collection $S$ of pages
  1. Relatively small
  2. Rich in relevant pages
  3. Contains most (or many) of the strongest authorities
- First (naive) solution
  - Only top $t$ pages containing the query
    - Not include the best authorities (condition 3)

Focused graph – Solution

- $t$ Top ranked pages ~ root set $R$
  - Subset of all pages containing the query
  - Extremely few links between pages in $R$
- Include $d$ pages pointing to $R$
- Include all pages linked from $R$
- $t = 200$, $d = 50$
  - $1000 < |R| < 5000$
Focused graph – Links

- Intrinsic
  - Links inside the domain
- Transverse
  - Links outside the domain
- Intrinsic links are removed from $S$

Focused graph – Goal

- Contains
  - Many relevant pages
  - Strong authorities
  - Many pages not relevant
- Goal
  - Extracting these authorities
Computing authorities and hubs

- Order pages by their in-degree
  - Rejected on the collection of all pages
  - Better on focused graph
- Problem
  - Query “java” – pages with high in-degree
    - www.gamelan.com, java.sun.com (Strong authorities)
    - vs.
    - www.amazon.com ("universally popular")

Problem – Observation

- Authorities
  - Overlap in the sets of pages that point to them (hubs)
- “Universally popular”
  - No significant overlap
Mutually reinforcing relationship

- Good hub
  - Points to many good authorities
- Good authority
  - Pointed by many good hubs

- Problem
  - Circularity

Iterative Algorithm

- Updates numerical weights for each page $p$
  - Squares sum normalized to 1
- Authority weight $x^{<p>}$
- Hub weight $y^{<p>}$
- $I$ operation
  - $x^{<p>} \leftarrow \sum_{q \in (p, q) \in E} y^{<q>}$
- $O$ operation
  - $y^{<p>} \leftarrow \sum_{q \in (p, q) \in E} x^{<q>}$
- Normalization

- Good hub
  - Points to many good authorities
- Good authority
  - Pointed by many good hubs
Iterative algorithm

Iterate(G, k)

G: a collection of n linked pages
k: a natural number
Let z denote the vector (1; 1; 1; ...; 1) ∈ ℝn.
Set x₀ := z:
Set y₀ := z:
For i = 1; 2; ...; k
  Apply the I operation to (x₀i−1; y₀i−1), obtaining new x-weights x₀i.
  Apply the O operation to (x₀i; y₀i−1), obtaining new y-weights y₀i.
  Normalize x₀i, obtaining xᵢ.
  Normalize y₀i, obtaining yᵢ.
End
Return (xₖ; yₖ).

Retrieving authorities and hubs

Filter(G, k, c)

G: a collection of n linked pages
k, c: natural numbers
(x₀; y₀) := Iterate(G; k).
Report the pages with the c largest coordinates in x₀ as authorities.
Report the pages with the c largest coordinates in y₀ as hubs.
Matrix, Eigenvalues, Eigenvectors

- $M^{m \times n}$ symmetric
- Eigenvalues $\lambda_1(M), \ldots, \lambda_n(M)$
  - In order of decreasing absolute values
  - Each listed a number of times equal to its multiplicity
- Eigenvectors $\omega_1(M), \ldots, \omega_n(M)$
- Assumption
  - $\lambda_1(M) > \lambda_2(M)$
- Principal eigenvector $\lambda_1(M)$
- Non-principal eigenvectors $\lambda_i(M), i = 2, \ldots, n$

Convergence

- Theorem: The sequences $x^i, \ldots$ and $y^i, \ldots$ converge (to limit $x^*$ and $y^*$ respectively).
- Theorem: $x^*$ and $y^*$ is the principal eigenvector of $ATA$ and $AAT$ respectively.

- Experiment
  - $C \approx 5-10$
    - $k = 20$
Iterative algorithm - Result

 iterative algorithm – Summary

- The text used only for retrieving the root set $R$, the following analysis ignored the text.
- The algorithm produces authorities with respect to the www as a whole, despite it operates without direct access to large-scale search engine.
- Analysis of the full www link structure can be replaced by an analysis on a small focused subgraph.
“What do users of the www consider to be related to p, when they create pages and hyperlink?”

If p highly referenced => Abundance Problem
  - Enormous number of independent opinions about the relation of p to others pages

Using authorities and hubs
  - “In the local region of the link structure near p, what are the strongest authorities?”

Iterative algorithm easily adopted
  - “Find t pages pointing to p.”
    - Instead of “Find t pages containing the query.”

• Ranking pages be their in-degrees is still not satisfactory.

http://www.honda.com Honda
http://www.ford.com/ Ford Motor Company
http://www.e.org/blueribbon.html The Blue Ribbon Campaign for Online Free Speech
http://www.mckinley.com/ Welcome to Magellan!
http://www.netscape.com Welcome to Netscape
http://www.linkexchange.com/ LinkExchange | Welcome
http://www.toyota.com/ Welcome to @Toyota
http://www.pointcom.com/ PointCom
http://home.netscape.com/ Welcome to Netscape
http://www.yahoo.com Yahoo!
Related Work– Social Network

- Standing
  - “importance” of individuals in an implicitly defined network
- Definitions
  - Edge (i, j) ~ “endorsement” of j by i
  - $A_{ij}$ ~ the strength/weight of the endorsement from i to j
  - $P^r_{ij}$ ~ number of paths of length exactly r from i to j
  - $b < 1$ constant small enough to $Q_g = \sum_{r=1}^{\infty} b^r P^r_{ij}$ convergences for each pair (i, j)
Related Work – Social Network

- $s_j$ standing of node $j$ (Katz)
  - $s_j = \sum_j Q_{ij}$
  - $I^{th}$ column of matrix $(I - bA)^{-t} - I$

- $s_j$ standing of node $j$ (Hubbell)
  - $s_j = e_j + \sum_i A_{ij} s_i$
  - $A_{ij}$ ~ strength of endorsement from $i$ to $j$
  - $e_j$ ~ estimate of the standing of node $j$
  - Standing of $j$ ~ total “quantity” of endorsement entering a node $j$, weighted by standing of endorsers
  - The vector of standings ~ $(I - AT)^{-1}e$

Related Work – Scientific Citations

- Importance and impart of individual scientific papers and journals
- Impact Factor of journal $j$ (Garfield)
  - Average number of citations received by papers published in the previous two years of journal $j$
  - Pure counting of the in-degrees
- Influential journal (Pinski, Narin)
  - Heavily cited by other influential journals
  - Parallel between this vs. hubs and authorities
Related Works – Scientific Citations

- \( A_{ij} \sim \) the fraction of citations from \( j \) to \( i \)
- Influence of \( j \) \( \sim w_j = \sum_i A_{ij}w_i \)
  - \( w \geq 0, w \neq 0, A^T w = w \)
  - \( w \) is a principal eigenvector of \( A^T \)
- Geller: Correspond to the stationary distribution of a random walk (without jumps)
- Solla Price, Noma: Handling journal self-citations

Related Works – Scientific Citations

- **Journals**
  - Highly authoritative journals on a common topic reference one another extensively
  - One-level model
- **Web**
  - The strongest authorities consciously do not link to one other
  - Two-level patterns (hubs and authorities)
Related Work – Hypertext and Rankings

- Index node
  - Its out-degree >> the average out-degree
- Reference node
  - Its in-degree >> the average in-degree

... missing ...
Multiple Sets of Hubs and Authorities

• There is no only the first eigenvector
• Other useful for queries
  o With several very different meaning ("jaguar")
  o Arise as a term in the context of multiple technical communities ("randomized algorithm")
  o Refer to a highly polarized issue ("abortion")
• The relevant documents can be naturally grouped into several clusters
Multiple Sets of Hubs and Authorities

(abortion) Authorities: 2nd non-principal vector, positive end
0.321 http://www.caral.org/abortion.html Abortion and Reproductive Rights
Internet Resources
0.219 http://www.plannedparenthood.org/ Welcome to Planned Parenthood
0.195 http://www.gynpages.com/ Abortion Clinics OnLine
0.172 http://www.oneworld.org/ippf/ IPPF Home Page
0.162 http://www.prochoice.org/naf/ The National Abortion Federation
0.161 http://www.lm.com/lmann/feminist/abortion.html

(abortion) Authorities: 2nd non-principal vector, negative end
0.197 http://www.awinc.com/partners/bc/commpass/lifenet/lifenet.htm LifeWEB
0.169 http://www.worldvillage.com/wv/square/chapel/xwalk/html/peter.htm Healing
after Abortion
0.164 http://www.nebula.net/maeve/lifelink.html
0.150 http://members.aol.com/pladvocate/ Pro-Life Advocate
0.144 http://www.clark.net/pub/jed/factbot.html The Right Side of the Web
0.144 http://www.catholic.net/HyperNews/get/abortion.html

Diffusion and Generalization

• Specific query
  ○ Not enough relevant pages in G
  ○ Authoritative pages of “broader” topic will win
• The process has diffused from the initial query.

("WWW conferences") Authorities: principal eigenvector
0.088 http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-
new.html The What’s New Archive
0.088 http://www.w3.org/hypertext/DataSources/WWW/Servers.html World-
Wide Web Servers: Summary
0.087 http://www.w3.org/hypertext/DataSources/bySubject/Overview.html
The World-Wide Web Virtual Library
Evaluation

- Iterative algorithm
  - Five top hubs + Five top authorities
- AltaVista
  - Top ten pages
- Yahoo! (managed list)
  - Used for comparison

Evaluation

- 26 search topics
- 37 users
- 1369 responses
  - “bad”, “fair”, “good”, “fantastics”

- (31%) Yahoo! and Iterative algorithm equivalent
- (50%) Iterative algorithm evaluated higher
- (19%) Yahoo! Evaluated higher
Summary

- The amount of relevant information on the www growing extremely rapidly
  - Way to distill a broad topic down to a representation of very small size ~ “authoritative” sources
- Producing results of a high quality
  - In the context of the www globally
- Infer global notions of structure without directly maintaining index of the www or its link structure

And now?

<table>
<thead>
<tr>
<th>Query</th>
<th>Google</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engines</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Search engine</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Engine search</td>
<td>-</td>
<td>OK</td>
</tr>
<tr>
<td>Harvard</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Java</td>
<td>OK</td>
<td>OK</td>
</tr>
</tbody>
</table>
It’s your turn now...

- Questions
- Suggestions
- Gifts
- Autographs

See you next week!

- Special thanks to
  - F. Ricci
  - Querists
- References are mentioned in the presented paper

Wake up the sleeping ones please. 😊