“Authoritative Sources in a Hyperlinked Environment”
by Jon M. Kleinberg

Paulius Miksys
Diana Zverelo
Contents

- The problem definition
- Solution (analysis of the link structure)
  - Constructing of a Focused Subgraph
  - Computing Hubs and Authorities
- Iterative algorithm
- Solution for similar-page queries
- Related work
- Multiple sets of hubs and authorities
- Diffusion and Generalization
- Evaluation
- Conclusions
Queries

- Specific queries. E.g., “Does Netscape support JDK 1.1 code-signing API?”
- Broad-topic queries. E.g., “Find information about the Java programming language.”
- Similar-page queries. E.g., “Find pages 'similar' to java.sun.com”
Problems accomplishing searches by queries

- For specific queries there is a *Scalarity Problem*: there are very few pages that contain the required information, and it is often difficult to determine the identity of these pages.

- For broad-topics queries exists *Abundance Problem*: The number of pages that could reasonably be returned as relevant is far too large for a human user to digest.

Two other complications:

- for a query “Harward”, [www.harward.edu](http://www.harward.edu) is one of the most authoritative pages. But there are millions of other pages that use “Harward” term and are higher in text-based searches.

- Natural authorities of the query “search engines” do not use this term on their pages as Honda or Toyota web-pages do not contain term “automobile manufacturers”.
Analysis of Link Structure as Solution

Why hyper links?

- Judgment of authority level (by creator of page $p$ when he makes a link to a page $q$ that supposed to be an authority)
- Potential authorities through pages that point to them

Pitfalls:

- Large number of links created for navigational purposes (“return to the main page”)
- Balance between relevance and popularity

Link-based model:

- Authorities
- Hubs (pages that link to many authorities)
Graph theory

- Directed graph \((G = (V, E))\)
- Out-degree of a node
- In-degree of a node
- Induced subgraph \((G[W])\)
Constructing of focused subgraph of WWW

We have a set created by text-based search engine.

Why do we need subset?

- the set may contain too many pages and entail a considerable computational cost
- most of the best authorities may not belong to this set

Subset properties:

- relatively small
- rich in relevant pages
- contains most ( or many ) of the strongest authorities
Subset construction

Subgraph(σ, E, t, d)
σ: a query string.
E: a text-based search engine.
t, d: natural numbers.
Let $R_σ$ denote the top t results of E on σ
Set $S_σ := R_σ$
For each page p ∈ $R_σ$
    Let $Γ^+(p)$ denote the set of all pages p points to.
    Let $Γ^-(p)$ denote the set of all pages pointing to p.
    Add all pages in $Γ^+(p)$ to $S_σ$.
    If $|Γ^-(p)| ≤ d$ then
        Add all pages in $Γ^-(p)$ to $S_σ$.
    Else
        Add an arbitrary set of d pages from $Γ^-(p)$ to $S_σ$.
End
Return $S_σ$
Subgraph reduction

- Offset the effect of links that serve purely a navigational function
  - remove all *intrinsic* edges from the graph, keeping only the edges corresponding to *transverse* links
  - Remove links that are mentioned in more than m pages (m=4-8).
Computing hubs and authorities

Ordering pages by their in-degree

**Pitfall:** this approach includes in $G_\sigma$ “universally popular” results which are not strong authorities

**Problem:** extract authorities from the latter constructed subgraph, excluding “universally popular” pages.

**Mutually reinforcing relationships:**
- 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
Hubs and Authorities

- Hubs
- Authorities

unrelated page of large in-degree
An Iterative Algorithm (I and O operation, weights)

- Authority weight: $x^{<p>}$
- Hub weight: $y^{<p>}$
- I operation: $x^{<p>} = \sum_{q: (q,p) \in E} y^{<p>}$
- O operation: $y^{<p>} = \sum_{q: (q,p) \in E} x^{<p>}$

sets of weights $x^{<p>}$ and $y^{<p>}$ are represented as vectors $x$ and $y$ of length 1
An Iterative Algorithm

Iterate($G,k$)

$G$: a collection of $n$ linked pages
$k$: a natural number

Let $z$ denote the vector $(1,1,1,...,1) \in \mathbb{R}^n$.

Set $x_0 := z$:
Set $y_0 := z$:

For $i = 1,2,...,k$
   Apply the $I$ operation to $(x_{i-1}, y_{i-1})$, obtaining new $x$-weights $x_{0i}$.
   Apply the $O$ operation to $(x_{0i}, y_{i-1})$, obtaining new $y$-weights $y_{0i}$.
   Normalize $x_{0i}$, obtaining $x_i$.
   Normalize $y_{0i}$, obtaining $y_i$.

End

Return $(x_k, y_k)$.

Filter($G,k,c$)

$G$: a collection of $n$ linked pages
$k,c$: natural numbers

$(x_k, y_k) := $Iterate($G; k$).

Report the pages with the $c$ largest coordinates in $x_k$ as authorities.
Report the pages with the $c$ largest coordinates in $y_k$ as hubs.
Eigenvectors and Eigenvalues

- M is symmetric n x n matrix
- **Eigenvalue** of M is a number $\lambda$ that $M\omega = \lambda\omega$
- Eigenvalues, indexed in order of decreasing absolute value: $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$.
- A set of **eigenvectors** $\omega_1(M), \omega_2(M), \ldots, \omega_n(M)$ that $\omega_i(M)$ belongs to the eigenspace of $\lambda_i(M)$
- Assumption $|\lambda_1(M)| > |\lambda_2(M)|$
- If assumption holds: $\omega_1(M)$ is a principal eigenvector, and $\omega_2(M), \ldots, \omega_n(M)$ are non-principal
Example:
The matrix $M$ has as two eigenvectors:

$v_1 = (1 \ 1)^t$ and $v_2 = (3 \ 1)^t$

$M \cdot v_1 = (-1 \ -1)^t = -1 \cdot v_1$

The eigenvalue is $-1$

$M \cdot v_2 = (3 \ 1)^t = 1 \cdot v_2$

The eigenvalue is $1$

\[
M = \begin{pmatrix}
2 & -3 \\
1 & -2
\end{pmatrix}
\]
Convergence (example)

\[ h = \lambda Aa \]
\[ a = \mu A^T h \]
\[ h = \lambda \mu AA^T h \]
\[ a = \lambda \mu A^TA a \]

An example is taken from
http://www.cs.uiowa.edu/~hzhang/c145/
Theorem: The sequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ converge (to limits $x^*$ and $y^*$ respectively)

Theorem: $x^*$ is the principal eigenvector of $A^TA$, and $y^*$ is the principal eigenvector of $AA^T$

$c=5-10$, $k = 20$
Results of broad-topic querying using in-degrees and using authorities

Ranking pages of $G_\sigma$ by their in-degrees when initial query is “java”:
- http://www.gamelan.com
- http://java.sun.com
- pages that propose Caribbean vacations
- Home page of Amazon Book

Top authorities obtained from $G_\sigma$ when initial query is “java”:
- http://www.gamelan.com
- http://java.sun.com
- http://lightyear.ncsa.uiuc.edu/~srp/java/javabooks.html
Method for broad-topic queries can be adapted to this situation with essentially no modification.

<table>
<thead>
<tr>
<th>Broad-topic queries</th>
<th>Similar-pages queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>query</td>
</tr>
<tr>
<td>String $\sigma$</td>
<td>Page $p$</td>
</tr>
<tr>
<td>“Find $t$ pages</td>
<td>“Find $t$ pages pointing</td>
</tr>
<tr>
<td>containing the string $\sigma$”</td>
<td>to $p$”</td>
</tr>
</tbody>
</table>

A root set $R_p$ consist $t$ pages that point to $p$; we grow it into a base set $S_p$ and result a subgraph $G_p$ in which we can search for hubs and authorities.
Solution for similar-page queries

Why linked-based analysis is better than text-based search for similar-pages queries

- Many of pages consists mostly of images
- Small quantity of text purely overlaps
- Proposed algorithm, is working on links, that creators of WWW pages tend to “classify” together with the given pages
Results of similar-page querying using in-degrees and using authorities

Ranking pages of \( G_p \) by their in-degrees when initial page is \( \text{www.honda.com} \):

- \( \text{www.honda.com} \)
- \( \text{www.ford.com} \)
- \( \text{www.eff.org/blueribbon.html} \)
- \( \text{www.mckinley.com} \)
- \( \text{www.netscape.com} \)
- \( \text{www.linkexchange.com} \)
- \( \text{www.toyota.com} \)
- \( \text{www.pointcom.com} \)

Top authorities obtained from \( G_p \) when initial page is \( \text{www.honda.com} \):

- \( \text{www.toyota.com} \)
- \( \text{www.honda.com} \)
- \( \text{www.ford.com} \)
- \( \text{www.bmwusa.com} \)
- \( \text{www.volvocars.com} \)
- \( \text{www.saturncars.com} \)
- \( \text{www.nissanmotors.com} \)
- \( \text{www.audi.com} \)
Related work

Definitions

- Standing - “importance” of individuals in an implicitly defined network.
- $G = (V, E)$ – graph of the network.
- Edge $(i, j)$ – “endorsement” of $j$ by $i$.
- $A$ – the matrix whose $(i, j)^{th}$ entry represents the strength of the endorsement from a node $i$ to $j$. 
Social networks

Models for counting standings:
- Katz
- Hubbell
Katz model

- Standing is based on the total number of paths terminating at node j, weighted by an exponentially decreasing damping factor.

\[ s_j = \sum_i Q_{ij} \]  
(standing of node j)

\[ Q_{ij} = \sum_{r=1}^{\infty} b^r P^{<r>}_{ij} \]

where

\[ P^{<r>}_{ij} \] - number of paths of length exactly r from i to j.

\[ b < 1 \] constant small enough to converge for each pair (i, j)
Hubbell model

- \( s_j = e_j + \sum_i A_{ij} s_i \)
- \( A_{ij} s_i \) - strength of endorsement from \( i \) to \( j \)
- \( e_j \) - estimate of the standing of node \( j \)

Standing is equal to the number of paths terminating at node \( j \), weighted by the standing of endorser, plus \( e_j \).
Scientific Citations - Garfield’s impact factor

- Citation-based measures of standing of journals

- Garfield’s impact factor
  - provides a numerical assessment of journal
  - the impact factor of a journal $j$ in a given year is the average number of citations received by papers published in the previous two years of journal $j$
  - based fundamentally on a pure counting of the in-degrees of nodes
Pinski and Narin

- Not all citations are equally important. A journal is “influential” if it is heavily cited by other influential journals.
- Parallel between this and hubs authorities
Hypertext and WWW rankings

- **Page-rank.**
  - Authority is passed directly from authorities to other authorities (no hub pages).
  - Applied to compute ranks for all the nodes of the www.
There can be found several densely linked collections of hubs and authorities among the same set $S_\sigma$ of pages. Each such collection could be well-separated from one another in the graph $G_\sigma$ for a variety of reasons:

- The query string may have several very different meanings. ("jaguar")
- The string may arise as a term in the context of multiple technical communities. ("randomized algorithms")
- String may refer to a highly polarized issue, involving groups that are not likely to link to one another. ("abortion")
Multiple Sets of Hubs and Authorities

- Principal eigenvector of $A^TA$ and $AA^T$ matrices is related to computed hubs and authorities.
- Non-principal eigenvectors of $A^TA$ and $AA^T$ can be used to extract additional densely linked collections of hubs and authorities.
Multiple Sets of Hubs and Authorities (Results)

- **(jaguar*) Authorities: principal eigenvector**
  - http://www2.ecst.csuchico.edu/jschlich/Jaguar/jaguar.html
  - http://www-und.ida.liu.se/t94patsa/jserver.html
  - http://tangram.informatik.uni-kl.de:8001/rgehm/jaguar.html
  - http://www.mcc.ac.uk/dlms/Consoles/jaguar.html

- **(jaguar jaguars) Authorities: 2nd non-principal vector, positive end**
  - http://www.jaguarsnfl.com/
  - http://www.nando.net/SportServer/football/nfl/jax.html
  - http://www.ao.net/brett/jaguar/index.html

- **(jaguar jaguars) Authorities: 3rd non-principal vector, positive end**
  - http://www.jaguarvehicles.com/
  - http://www.collection.co.uk/
  - http://www.coys.co.uk/
Multiple Sets of Hubs and Authorities (Results)

- (“randomized algorithms”) Authorities: 1st non-principal vector, positive end
  - http://theory.lcs.mit.edu/goemans/
  - http://theory.lcs.mit.edu/spielman/
  - http://www.nada.kth.se/johanh/
  - http://theory.lcs.mit.edu/rivest/

- (“randomized algorithms”) Authorities 1st non-principal vector, negative end
  - -.00116 http://lib.stat.cmu.edu/
  - http://www.geo.fmi./prog/tela.html
  - http://gams.nist.gov/
  - http://www.netlib.org

- (“randomized algorithms”) Authorities 4th non-principal vector, negative end
  - http://www.amara.com/current/wavelet.html
  - http://www.ocean.tamu.edu/baum/wavelets.html
  - http://www.mathsoft.com/wavelets.html
  - http://www.mat.sbg.ac.at/uhl/wav.html
Diffusion and Generalization

- **Specific queries**
  - Scaliarity problem - not enough relevant pages in $G_q$
  - “broader” topics win out over the pages relevant to $q$

- The process diffuses from the initial query
Diffusion and generalization (example)

- **(WWW conferences")** Authorities: principal eigenvector
  - http://www.w3.org/hypertext/DataSources/WWW/Servers.html World-Wide Web Servers: Summary
  - http://www.w3.org/hypertext/DataSources/bySubject/Overview.html The World-Wide Web Virtual Library

- **While text-based search engine produces 300 of pages containing the string**

- **(WWW conferences")** Authorities: 11th non-principal vector, negative end
  - http://www.csu.edu.au/special/conference/WWWWWW.html AUUG'95 and Asia-Pacfic WWW'95 Conference
Evaluation

- Attempting to define and compute “authority”, that is inherently based on human judgment
- Examples of output showed in article:
  - To show type of results that are produced
  - Res ipsa loquitur ("the thing itself speaks" in Latin)
- To evaluate an algorithm
  - 26 search topics were used as queries in search engines
  - 37 users that were not experts in CS and in 26 analyzed topics
  - 1369 responses used to assess the relative quality of Iterative algorithm, Yahoo! And Altavista on each topic
- Result
  - (31%) Yahoo! and Iterative algorithm equivalent
  - (50%) Iterative algorithm evaluated higher
  - (19%) Yahoo! Evaluated higher
Summary and Conclusions

- The amount of relevant information is growing extremely rapidly. Find the way to distill a broad topic down to a representation of very small size of “authoritative” sources.

- Produce results that are of as a high a quality as possible in the context of what is available on the www globally.

- Infer global notions of structure without directly maintaining index of the www or its link structure.