

---

# “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 *Scalability 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 to 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( $\sigma$ , E, t, d)**

$\sigma$ : a query string.

E: a text-based search engine.

t, d: natural numbers.

Let  $R_\sigma$  denote the top t results of E on  $\sigma$

Set  $S_\sigma := R_\sigma$

For each page  $p \in R_\sigma$

Let  $\Gamma^+(p)$  denote the set of all pages p points to.

Let  $\Gamma^-(p)$  denote the set of all pages pointing to p.

Add all pages in  $\Gamma^+(p)$  to  $S_\sigma$ .

If  $|\Gamma^-(p)| \leq d$  then

Add all pages in  $\Gamma^-(p)$  to  $S_\sigma$ .

Else

Add an arbitrary set of d pages from  $\Gamma^-(p)$  to  $S_\sigma$ .

End

Return  $S_\sigma$

---

# 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

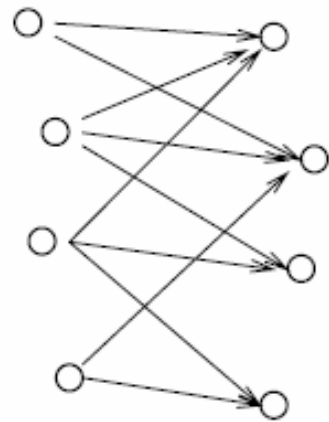
**Pitfall:** this approach includes in  $G_\sigma$  “universally popular” results which are not strong authorities

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

## Mutually reinforcing relationships:

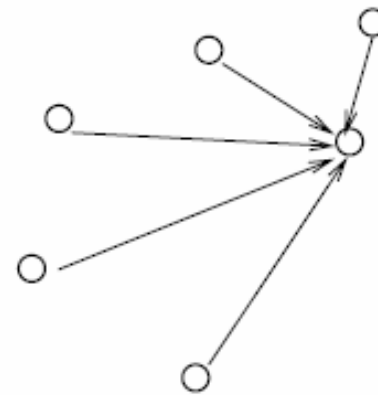
- a good *hub* is a page that points to many good authorities
  - a good *authority* is a page that is pointed to by many good hubs
-

# 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  $(1, 1, 1, \dots, 1) \in \mathbf{R}^n$ .

Set  $x_0 := z$ :

Set  $y_0 := z$ :

For  $i = 1, 2, \dots, k$

    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

$(x_k; y_k) := \text{Iterate}(G; k)$ .

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  $n \times n$  matrix
- **Eigenvalue** of  $M$  is a number  $\lambda$  that  $M\omega = \lambda\omega$
- Eigenvalues, indexed in order of decreasing absolute value:  $\lambda_1(M), \lambda_2(M), \dots, \lambda_n(M)$ .
- A set of **eigenvectors**  $\omega_1(M), \omega_2(M), \dots, \omega_n(M)$  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  $\omega_2(M), \dots, \omega_n(M)$  are *non-principal*

# Eigenvectors and Eigenvalues (example)

- Example:

The matrix  $M$  has as two eigenvectors:

$$\mathbf{v1} = (1 \ 1)^t \text{ and } \mathbf{v2} = (3 \ 1)^t$$

$$\mathbf{M}^*\mathbf{v1} = (-1 \ -1)^t = -1 \ \mathbf{v1}$$

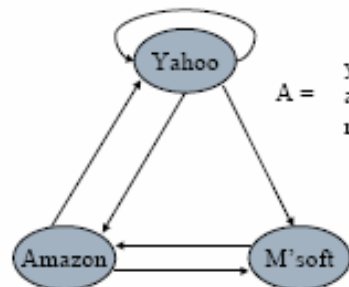
The eigenvalue is -1

$$\mathbf{M}^*\mathbf{v2} = (3 \ 1)^t = 1 \ \mathbf{v2}$$

The eigenvalue is 1

$$M = \begin{pmatrix} 2 & -3 \\ 1 & -2 \end{pmatrix}$$

# Convergence (example)



$$A = \begin{array}{c|ccc} & y & a & m \\ \hline y & 1 & 1 & 1 \\ a & 1 & 0 & 1 \\ m & 0 & 1 & 0 \end{array}$$

$$A^T = \begin{array}{ccc|c} 1 & 1 & 0 & \\ \hline 1 & 0 & 1 & \\ 1 & 1 & 0 & \end{array}$$

a(yahoo)	=	1	1	1	1	...	1
a(amazon)	=	1	1	4/5	0.75	...	0.732
a(m'soft)	=	1	1	1	1	...	1
h(yahoo)	=	1	1	1	1	...	1.000
h(amazon)	=	1	2/3	0.71	0.73	...	0.732
h(m'soft)	=	1	1/3	0.29	0.27	...	0.268

$$\mathbf{h} = \lambda A \mathbf{a}$$

$$\mathbf{a} = \mu A^T \mathbf{h}$$

$$\mathbf{h} = \lambda \mu A A^T \mathbf{h}$$

$$\mathbf{a} = \lambda \mu A^T A \mathbf{a}$$

An example is taken from

<http://www.cs.uiowa.edu/~hzhang/c145/>

# Convergence

- Theorem: The sequences  $x_1, x_2, \dots$  and  $y_1, y_2, \dots$  converge (to limits  $x^*$  and  $y^*$  respectively)
- Theorem:  $x^*$  is the principal eigenvector of  $A^T A$ , and  $y^*$  is the principal eigenvector of  $A A^T$
- $c=5-10, k = 20$

# 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://www.digitalfocus.com/digitalfocus/faq/howdoi.html>
- ❑ <http://lightyear.ncsa.uiuc.edu/~srp/java/javabooks.html>
- ❑ <http://ava.sun.com/aboutJava/index.html>

# Solution for similar-page queries

Method for broad-topic queries can be adapted to this situation with essentially no modification

	Broad-topic queries	Similar-pages queries
query	String $\sigma$	Page $p$
search	“Find $t$ pages containing the string $\sigma$ ”	“Find $t$ pages pointing to $p$ ”

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 [www.honda.com](http://www.honda.com):

- ❑ [www.honda.com](http://www.honda.com)
- ❑ [www.ford.com](http://www.ford.com)
- ❑ [www.eff.org/blueribbon.html](http://www.eff.org/blueribbon.html)
- ❑ [www.mckinley.com](http://www.mckinley.com)
- ❑ [www.netscape.com](http://www.netscape.com)
- ❑ [www.linkexchange.com](http://www.linkexchange.com)
- ❑ [www.toyota.com](http://www.toyota.com)
- ❑ [www.pointcom.com](http://www.pointcom.com)

Top authorities obtained from  $G_p$  when initial page is [www.honda.com](http://www.honda.com):

- ❑ [www.toyota.com](http://www.toyota.com)
- ❑ [www.honda.com](http://www.honda.com)
- ❑ [www.ford.com](http://www.ford.com)
- ❑ [www.bmwusa.com](http://www.bmwusa.com)
- ❑ [www.volvocars.com](http://www.volvocars.com)
- ❑ [www.saturncars.com](http://www.saturncars.com)
- ❑ [www.nissanmotors.com](http://www.nissanmotors.com)
- ❑ [www.audi.com](http://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  $(i, j)^{\text{th}}$  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 converges 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
-

---

# Scientific Citations – Pinski and Narin

## ■ Pinski and Narin

- ❑ Not all citations are equally important. A journal is “influential” if 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



---

# Multiple Sets of Hubs and Authorities

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  $A^T A$  and  $AA^T$  matrices is related to computed hubs and authorities.
  - Non-principal eigenvectors of  $A^T A$  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/rgehml/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>
    - <http://www.usatoday.com/sports/football/sfn/sfn30.htm>
  - **(jaguar jaguars) Authorities: 3rd non-principal vector, positive end**
    - <http://www.jaguarvehicles.com/>
    - <http://www.collection.co.uk/>
    - <http://www.moran.com/sterling/sterling.html>
    - <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

- Scalarity 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.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-new.html> The What's New Archive
    - <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.igd.fhg.de/www95.html> Third International World-Wide Web Conference
    - <http://www.csu.edu.au/special/conference/WWWWW.html> AUUG'95 and Asia-Pacic WWW'95 Conference
    - <http://www.ncsa.uiuc.edu/SDG/IT94/IT94Info.html> The Second International WWW Conference '94
    - <http://www.w3.org/hypertext/Conferences/WWW4/> Fourth International World Wide Web Conference
    - <http://www.igd.fhg.de/www/www95/papers/> WWW'95: Papers
-

---

# 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
-