

# Authoritative Sources in a Hyperlinked Environment

1

**JON M. KLEINBERG (PAPER)**  
**MARTIN CETKOVSKÝ (PRESENTATION)**

<http://martin.alikuvkoutek.cz>

Martin Cetkovský, JMK: ASHE

24 October 2006

## Table of contents

2

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

Martin Cetkovský, JMK: ASHE

24 October 2006

## Introduction

3

- *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

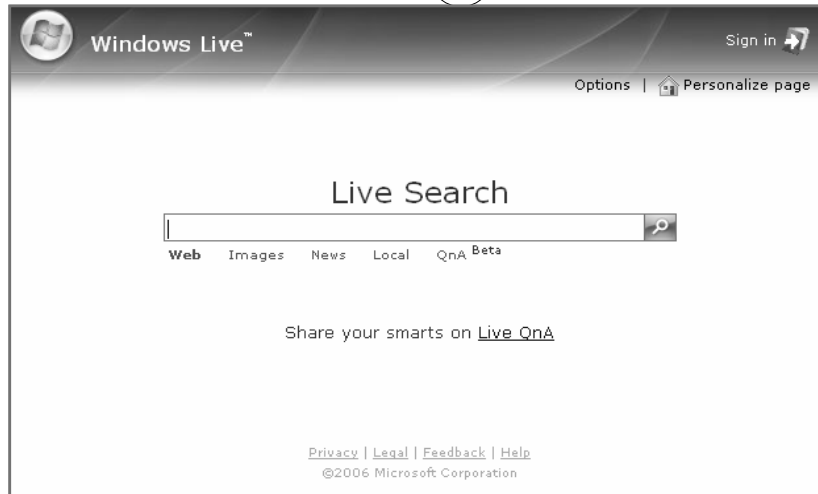
## Problem – Query types

4

- 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

5



## Hyperlink

6

- 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”

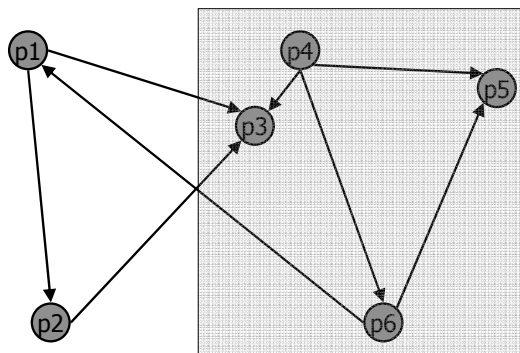
## Graph theory

7

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

## Directed graph – Example

8



## Graph model

9

- **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

10

- **Authorities**
  - Most important pages related to the query
  - Linked by many hubs
- **Hubs**
  - Pages that link to many related authorities

## Focused graph

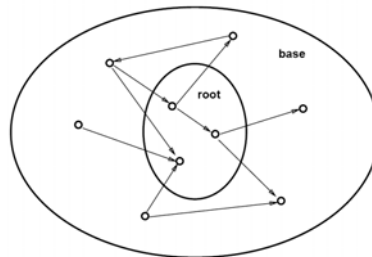
11

- 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

12

- $t$  Top ranked pages  $\sim$  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

13

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

## Focused graph – Goal

14

- **Contains**
  - Many relevant pages
  - Strong authorities
  - Many pages not relevant
- **Goal**
  - Extracting these authorities

## Computing authorities and hubs

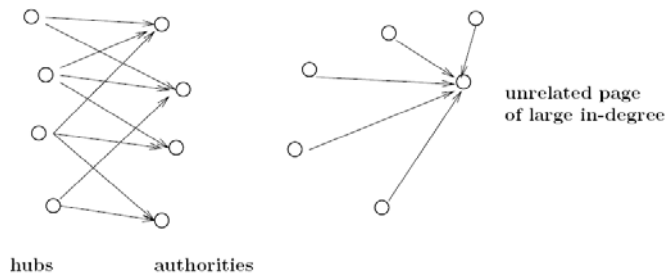
15

- 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

16

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



## Mutually reinforcing relationship

17

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

## Iterative Algorithm

18

- Updates numerical weights for each page  $p$ 
    - Squares sum normalized to 1
  - Authority weight  $x^{<p>}$
  - Hub weight  $y^{<p>}$
  - $I$  operation
    - $x^{<p>} \leftarrow \sum_{q:(q,p) \in E} y^{<q>}$
  - $O$  operation
    - $y^{<p>} \leftarrow \sum_{q:(p,q) \in E} x^{<q>}$
  - Normalization
- Good hub
    - Points to many good authorities
  - Good authority
    - Pointed by many good hubs

## Iterative algorithm

19

```
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  $(x_k; y_k)$ .
```

## Retrieving authorities and hubs

20

```
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.
```

## Matrix, Eigenvalues, Eigenvectors

21

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

## Convergence

22

- Theorem: The sequences  $x^t, \dots$  and  $y^t, \dots$  converge (to limit  $x^*$  and  $y^*$  respectively).
- Theorem:  $x^*$  and  $y^*$  is the principal eigenvector of  $A^T A$  and  $A A^T$  respectively.
- Experiment
  - $C \approx 5-10$
  - ✖  $k = 20$

## Iterative algorithm - Result

23

```
(java) Authorities
.328 http://www.gamelan.com/ Gamelan
.251 http://java.sun.com/ JavaSoft Home Page
.190 http://www.digitalfocus.com/digitalfocus/faq/howdoi.html The
  Java Developer: How Do I...
.190 http://lightyear.ncsa.uiuc.edu/srp/java/javabooks.html The Java
  Book Pages
.183 http://sunsite.unc.edu/javafaq/javafaq.html comp.lang.java FAQ

("search engines") Authorities
.346 http://www.yahoo.com/ Yahoo!
.291 http://www.excite.com/ Excite
.239 http://www.mckinley.com/ Welcome to Magellan!
.231 http://www.lycos.com/ Lycos Home Page
.231 http://www.altavista.digital.com/ AltaVista: Main Page
```

## Iterative algorithm – Summary

24

- 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.

## Similar-Page Queries

25

- “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*.”

## Similar-Page Queries

26

- Ranking pages by 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!*

## Similar-Page Queries – Result

27

(www.honda.com) Authorities  
.202 http://www.toyota.com/ *Welcome to @Toyota*  
.199 http://www.honda.com/ *Honda*  
.192 http://www.ford.com/ *Ford Motor Company*  
.173 http://www.bmwusa.com/ *BMW of North America, Inc.*  
.162 http://www.volvocars.com/ *VOLVO*  
.158 http://www.saturncars.com/ *Welcome to the Saturn Web Site*  
.155 http://www.nissanmotors.com/ *NISSAN - ENJOY THE RIDE*  
.145 http://www.audi.com/ *Audi Homepage*  
.139 http://www.4adodge.com/ *1997 Dodge Site*  
.136 http://www.chryslercars.com/ *Welcome to Chrysler*

## Related Work– Social Network

28

- **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
  - $P_{ij}^{<r>}$  ~ number of paths of length exactly r from i to j
  - $b < 1$  constant small enough to  $Q_{ij} = \sum_{r=1}^{\infty} b^r P_{ij}^{<r>}$  convergences for each pair (i, j)

## Related Work – Social Network

29

- $s_j$  standing of node  $j$  (Katz)
  - $s_j = \sum_i Q_{ij}$
  - $j^{\text{th}}$  column of matrix  $(I - bA)^{-1} - I$
- $s_j$  standing of node  $j$  (Hubbell)
  - $s_j = e_j + \sum_i A_{ij}s_i$
  - $A_{ij} \sim$  strength of endorsement from  $i$  to  $j$
  - $e_j \sim$  estimate of the standing of node  $j$
  - Standing of  $j \sim$  total “quantity” of endorsement entering a node  $j$ , weighted by standing of endorsers
  - The vector of standings  $\sim (I - A^T)^{-1}e$

## Related Work – Scientific Citations

30

- 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

31

- $A_{ij} \sim$  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

32

- 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

33

- Index node
  - Its out-degree  $\gg$  the average out-degree
- Reference node
  - Its in-degree  $\gg$  the average in-degree

34

- ... missing ...

## Multiple Sets of Hubs and Authorities

35

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

## Multiple Sets of Hubs and Authorities

36

("randomized algorithms") Authorities: 1st non-principal vector, positive end  
.125 <http://theory.lcs.mit.edu/goemans/> *Michel X. Goemans*  
.122 <http://theory.lcs.mit.edu/spielman/> *Dan Spielman's Homepage*  
.122 <http://www.nada.kth.se/johanh/> *Johan Hastad*  
.122 <http://theory.lcs.mit.edu/rivest/> *Ronald L. Rivest : HomePage*

("randomized algorithms") Authorities 1st non-principal vector, negative end  
.00116 <http://lib.stat.cmu.edu/> *StatLib Index*  
.00115 <http://www.geo.fmi./prog/tela.html> *Tela*  
.00107 <http://gams.nist.gov/> *GAMS : Guide to Available Mathematical Software*  
.00107 <http://www.netlib.org> *Netlib*

("randomized algorithms") Authorities 4th non-principal vector, negative end  
.176 <http://www.amara.com/current/wavelet.html> *Amara's Wavelet Page*  
.172 <http://www-ocean.tamu.edu/baum/wavelets.html> *Wavelet sources*  
.161 <http://www.mathsoft.com/wavelets.html> *Wavelet Resources*  
.143 <http://www.mat.sbg.ac.at/uhl/wav.html> *Wavelets*

## Multiple Sets of Hubs and Authorities

37

(abortion) Authorities: 2nd non-principal vector, positive end  
.321 <http://www.caral.org/abortion.html> *Abortion and Reproductive Rights Internet Resources*  
.219 <http://www.plannedparenthood.org/> *Welcome to Planned Parenthood*  
.195 <http://www.gynpages.com/> *Abortion Clinics OnLine*  
.172 <http://www.oneworld.org/ippf/> *IPPF Home Page*  
.162 <http://www.prochoice.org/naf/> *The National Abortion Federation*  
.161 <http://www.lm.com/lmann/feminist/abortion.html>

(abortion) Authorities: 2nd non-principal vector, negative end  
.197 <http://www.awinc.com/partners/bc/compass/lifenet/lifenet.htm> *LifeWEB*  
.169 <http://www.worldvillage.com/wv/square/chapel/xwalk/html/peter.htm> *Healing after Abortion*  
.164 <http://www.nebula.net/maeve/lifelink.html>  
.150 <http://members.aol.com/pladvocate/> *Pro-Life Advocate*  
.144 <http://www.clark.net/pub/jed/factbot.html> *The Right Side of the Web*  
.144 <http://www.catholic.net/HyperNews/get/abortion.html>

## Diffusion and Generalization

38

- 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  
.088 <http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-new.html> *The What's New Archive*  
.088 <http://www.w3.org/hypertext/DataSources/WWW/Servers.html> *World-Wide Web Servers: Summary*  
.087 <http://www.w3.org/hypertext/DataSources/bySubject/Overview.html> *The World-Wide Web Virtual Library*

## Evaluation

39

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

## Evaluation

40

- 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

41

- 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?

42

Query	Google	Yahoo
Search engines	-	?
Search engine	OK	OK
Engine search	-	OK
Harvard	OK	OK
Java	OK	OK

## It's your turn now...

43

- Questions
- Suggestions
  
- Gifts
- Autographs

## See you next week!

44

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

Wake up the sleeping ones please. ☺