Information Retrieval

Weike Pan

The slides are **adapted from those provided by Prof. Hinrich Schütze** at University of Munich (http://www.cis.lmu.de/~hs/teach/14s/ir/).

Chapter 21 Link analysis

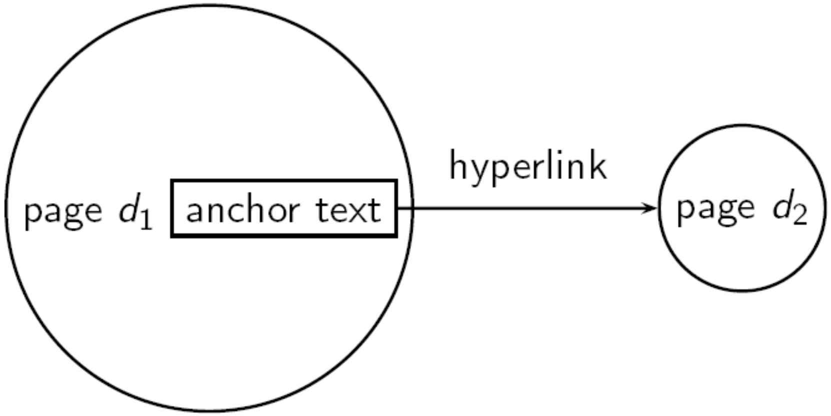
- 21.1 The Web as a graph
- 21.2 PageRank
- 21.3 Hubs and Authorities
- 21.4 References and further reading

Outline

- 21.1 The Web as a graph
- 21.2 PageRank
- 21.3 Hubs and Authorities
- 21.4 References and further reading

21.1 The Web as a graph

- The web as a directed graph

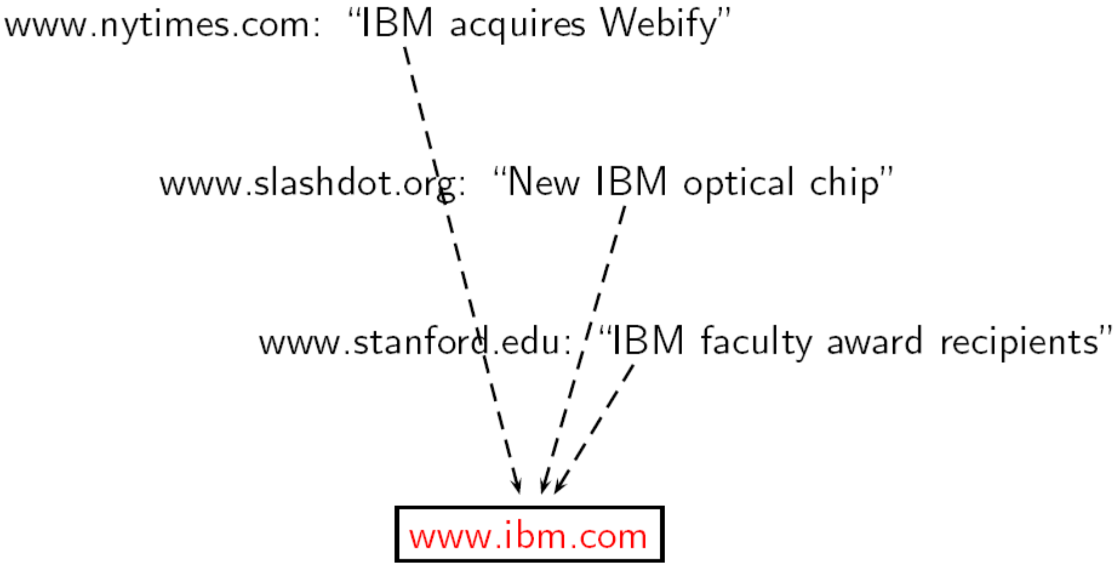


- Assumption 1: A hyperlink is a quality signal.
  - The hyperlink d1 -> d2 indicates that d1's author deems (认为) d2 high-quality and relevant.
- Assumption 2: The anchor text describes the content of d2.
  - We use anchor text somewhat loosely here for the text surrounding the hyperlink.

21.1 The Web as a graph

- Searching on [text of d2] + [anchor text -> d2] is often more effective than searching on [text of d2] only.

- Searching on [anchor text -> d2] is better for the query *IBM*.
  - In this representation, the page with the most occurrences of *IBM* is www.ibm.com.



21.1 The Web as a graph

- Indexing anchor text
  - Anchor text is often a better description of a page's content than the page itself.
  - Anchor text can be weighted more highly than the document text (based on Assumptions 1 and 2).

21.1 The Web as a graph

- Question
  - Assumption 1: A link on the web is a quality signal – the author of the link thinks that the linked-to page is of high quality.
    - Is assumption 1 true in general?
  - Assumption 2: The anchor text describes the content of the linked-to page.
    - Is assumption 2 true in general?

21.1 The Web as a graph

- The terms Google bombing and Googlewashing refer to the practice of causing a website to rank highly in web search engine results for irrelevant, unrelated or off-topic search terms by linking heavily.

https://en.wikipedia.org/wiki/Google_bomb

21.1 The Web as a graph

- **Citation analysis**: analysis of citations in the scientific literature
- Example citation: “Miller (2001) has shown that physical activity …”
  - We can view “Miller (2001)” as a hyperlink linking two scientific articles (即本论文和Miller (2001)论文).
- One application of these “hyperlinks” in the scientific literature:
  - Measure the similarity of two articles by the overlap of other articles citing them. This is called co-citation similarity.

21.1 The Web as a graph

- Another application: Citation frequency can be used to measure the impact of a scientific article
  - Simplest measure: Each citation gets one vote, citation frequency = inlink count

- However: A high inlink count does not necessarily mean high quality… mainly because of link spam.
  - Better measure: weighted citation frequency or citation rank
    - This is basically PageRank, which was invented in the context of citation analysis.

Outline

- 21.1 The Web as a graph
- 21.2 PageRank
- 21.3 Hubs and Authorities
- 21.4 References and further reading

21.2 PageRank

- Imagine a web surfer doing a random walk on the web
  - Start at a random page
  - At each step, go out of the current page along one of the links on that page, equiprobably (相同概率地)

- In the steady state, each page has a long-term visit rate

- This long-term visit rate is the page’s PageRank

- PageRank = long-term visit rate = steady state probability

21.2 PageRank

- **Formalization of random walk: Markov chains**

  - A Markov chain consists of N states, plus an $N \times N$ transition probability matrix P.
  - state = page
  - At each step, we are on exactly one of the pages.
  - For $1 \leq i, j \leq N$, the matrix entry P_{ij} tells us the probability of j being the next page, given we are currently on page i.
  - Clearly, for each i, $\sum_{j=1}^{N} P_{ij} = 1$

$$d_i \xrightarrow{P_{ij}} d_j$$

21.2 PageRank

- Long-term visit rate of page *d* is the probability that a web surfer is at page *d* at a given point in time.

- What properties must hold of the web graph for the long-term visit rate to be well defined?

- The web graph must correspond to an ergodic Markov chain
  - Irreducibility (不可约): There is a path from any page to any other page.
  - Aperiodicity (非周期): The pages cannot be partitioned such that the random walker visits the partitions sequentially.

21.2 PageRank

- At a dead end, jump to a random web page with probability 1/N.

- At a non-dead end
  - With probability 10%, jump to a random web page (to each with a probability of 0.1/N)
  - With remaining probability 90%, go out on a random hyperlink
  - 10% is a parameter called the teleportation rate

- Note: “jumping” from a dead end is independent of the teleportation rate.

- With teleporting, we cannot get stuck in a dead end.
- Teleporting makes the web graph ergodic.

21.2 PageRank

- **Calculation of PageRank (1/2)**
  - $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ is the PageRank vector, i.e., the vector of steady-state probabilities
  - If the distribution in this step is $\vec{x}$ (probability vector), then the distribution in the next step is $\vec{x}P$
  - Because $\vec{\pi}$ is the steady state, we have $\vec{\pi} = \vec{\pi}P$
  - Solving this matrix equation gives us $\vec{\pi}$, which is the principal left eigenvector for P, i.e., $\vec{\pi}$ is the left eigenvector with the largest eigenvalue
  - All transition probability matrices have largest eigenvalue 1

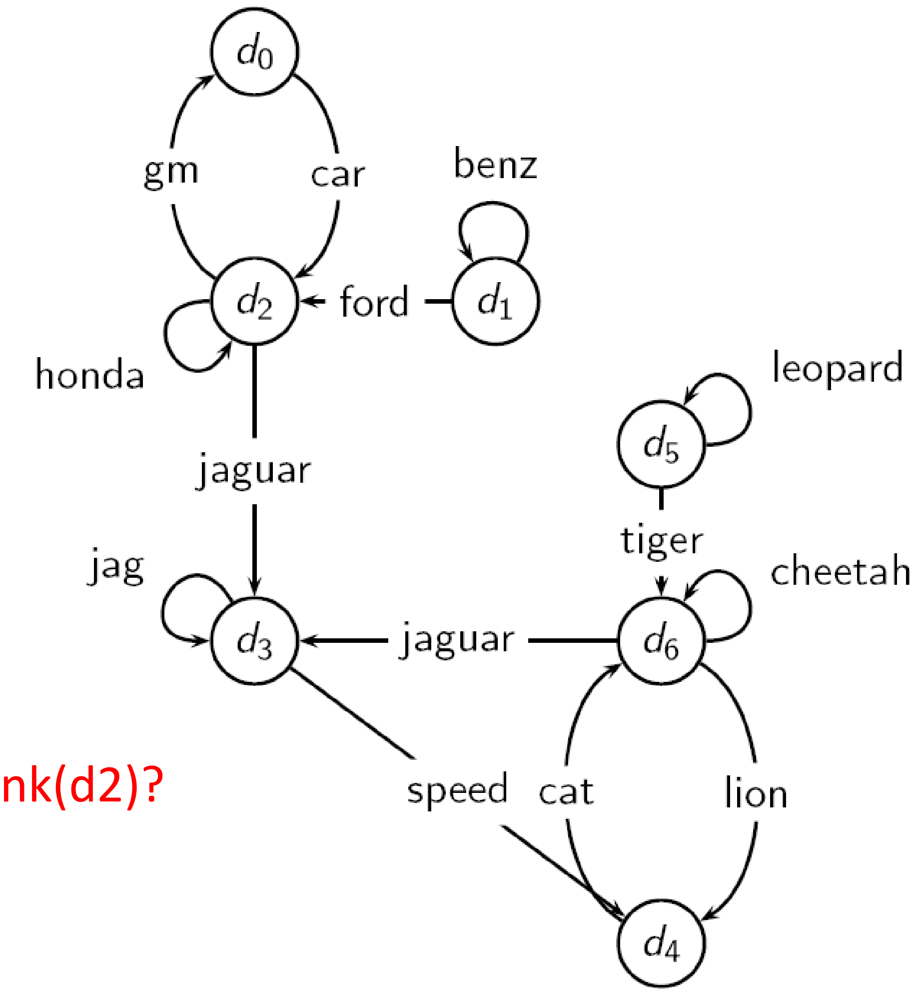
21.2 PageRank

- **Calculation of PageRank (2/2)**
  - Start with any distribution $\vec{x}$, e.g., uniform distribution
  - After one step, we're at $\vec{x}P$.
  - After two steps, we're at $\vec{x}P^2$.
  - After k steps, we're at $\vec{x}P^k$.
  - Algorithm: multiply $\vec{x}$ by increasing powers of P until convergence.
  - This is called the power method.
  - Regardless of where we start, we eventually reach the steady state $\vec{\pi}$

21.2 PageRank

- **Example web graph**

	PageRank
d_0	0.05
d_1	0.04
d_2	0.11
d_3	0.25
d_4	0.21
d_5	0.04
d_6	0.31



Why PageRank(d6) > PageRank(d2)?

21.2 PageRank

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Step 1. Link matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Step 2. Transition probability matrix

P

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.02	0.02	0.88	0.02	0.02	0.02	0.02
d_1	0.02	0.45	0.45	0.02	0.02	0.02	0.02
d_2	0.31	0.02	0.31	0.31	0.02	0.02	0.02
d_3	0.02	0.02	0.02	0.45	0.45	0.02	0.02
d_4	0.02	0.02	0.02	0.02	0.02	0.02	0.88
d_5	0.02	0.02	0.02	0.02	0.02	0.45	0.45
d_6	0.02	0.02	0.02	0.31	0.31	0.02	0.31

Step 3. Transition matrix with teleporting

	$\vec{x}$	$\vec{x}P^1$	$\vec{x}P^2$	$\vec{x}P^3$	$\vec{x}P^4$	$\vec{x}P^5$	$\vec{x}P^6$	$\vec{x}P^7$	$\vec{x}P^8$	$\vec{x}P^9$	$\vec{x}P^{10}$	$\vec{x}P^{11}$	$\vec{x}P^{12}$	$\vec{x}P^{13}$
d_0	0.14	0.06	0.09	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
d_1	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_2	0.14	0.25	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11
d_3	0.14	0.16	0.23	0.24	0.24	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25
d_4	0.14	0.12	0.16	0.19	0.19	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
d_5	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_6	0.14	0.25	0.23	0.25	0.27	0.28	0.29	0.29	0.30	0.30	0.30	0.30	0.31	0.31

Step 4. Power method

21.2 PageRank

- **Application of PageRank in IR**
  - Step 1: Query processing
  - Step 2: Retrieve pages satisfying the query
  - Step 3: Rank them by their PageRank (In practice: rank according to weighted combination of raw text match, anchor text match, PageRank and other factors)
  - Step 4: Return a re-ranked list to the user

21.2 PageRank

- **How important is PageRank?**
  - Frequent claim: PageRank is the most important component of web ranking.
- **The reality:**
  - There are several components that are at least as important, e.g., anchor text, phrases, proximity, tiered indexes ...
  - Rumor has it that PageRank in its original form (as presented here) now has a negligible impact on ranking!
  - However, variants of a page’s PageRank are still an essential part of ranking.
  - Addressing link spam is difficult and crucial.

Outline

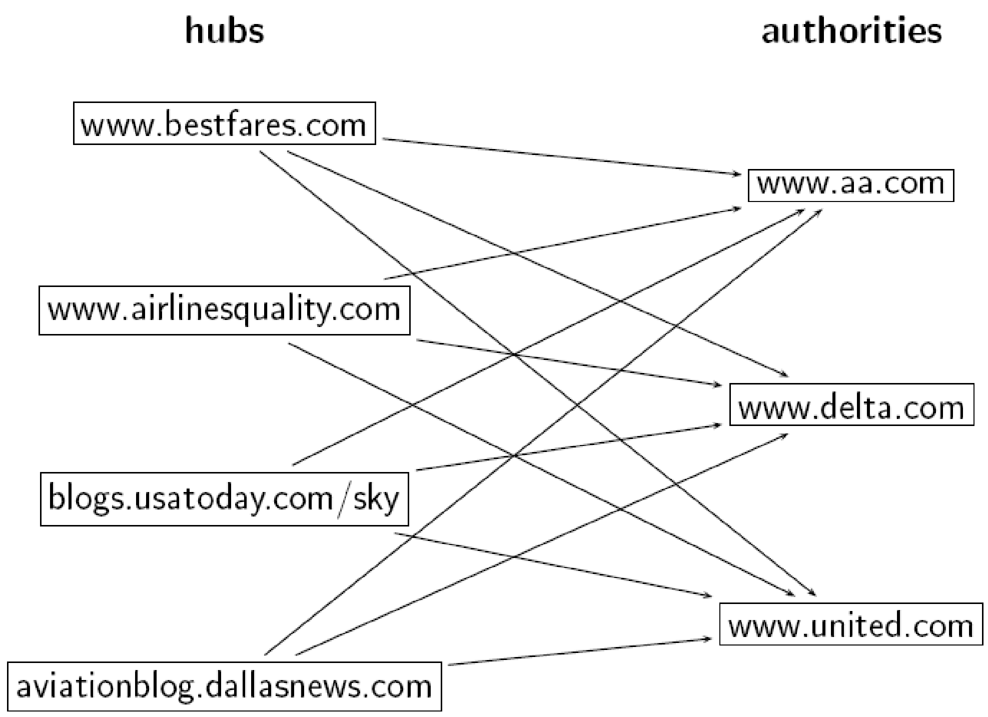
- 21.1 The Web as a graph
- 21.2 PageRank
- 21.3 Hubs and Authorities
- 21.4 References and further reading

21.3 Hubs and Authorities

- There are two different types of relevance on the web
- Relevance type 1: Hubs. A hub page is a good list of [links to pages answering the information need].
- Relevance type 2: Authorities. An authority page is a direct answer to the information need.
- Most approaches to search (including PageRank ranking) **don't make the distinction** between these two very different types of relevance.

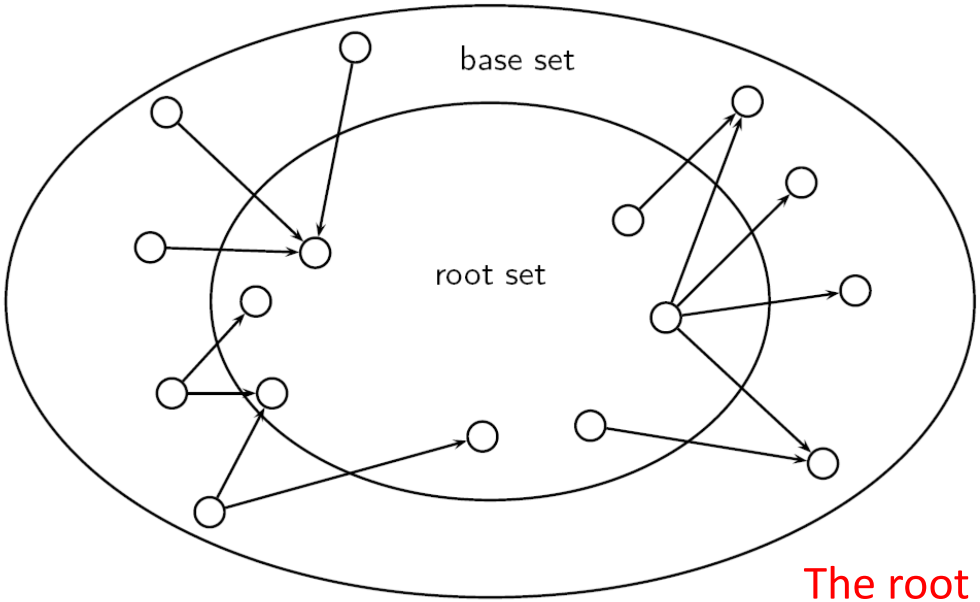
21.3 Hubs and Authorities

- A good hub page for a topic links to many authority pages for that topic.
- A good authority page for a topic is linked to by many hub pages for that topic.
- **Circular definition** -- we will turn this into an **iterative computation**.



21.3 Hubs and Authorities

- Do a regular web search first. Call the search result the **root set**.
- Find all pages that are linked from or link to pages in the root set. Call this larger set the **base set**.
- Finally, compute hubs and authorities for the **base set** (which we'll view as a small web graph)



The root set is a **subset** of the base set.

21.3 Hubs and Authorities

- Root set typically has 200-1000 nodes
- Base set may have up to 5000 nodes

- Computation of base set, as shown on the previous slide
  - Follow **outlinks** by parsing the pages in the root set
  - Find d’s **inlinks** by searching for all pages containing a link to d

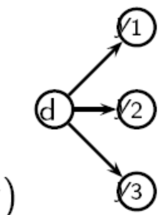
21.3 Hubs and Authorities

- HITS can pull together good pages regardless of page content.
- Once the base set is assembled, we only do link analysis, no text matching.
- Pages in the base set often do not contain any of the query words.
- In theory, an English query can retrieve Japanese-language pages if supported by the link structure between English and Japanese pages.
- Danger: topic drift – the pages found by following links may not be related to the original query.

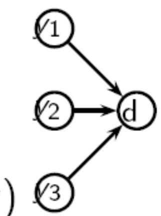
21.3 Hubs and Authorities

- Compute for each page d in the **base set** a hub score h(d) and an authority score a(d)

- Initialization: for all d: h(d) = 1, a(d) = 1
- **Iteratively update all h(d), a(d)**



- For all d: $h(d) = \sum_{d \mapsto y} a(y)$



- For all d: $a(d) = \sum_{y \mapsto d} h(y)$
- Iterate these two steps until convergence

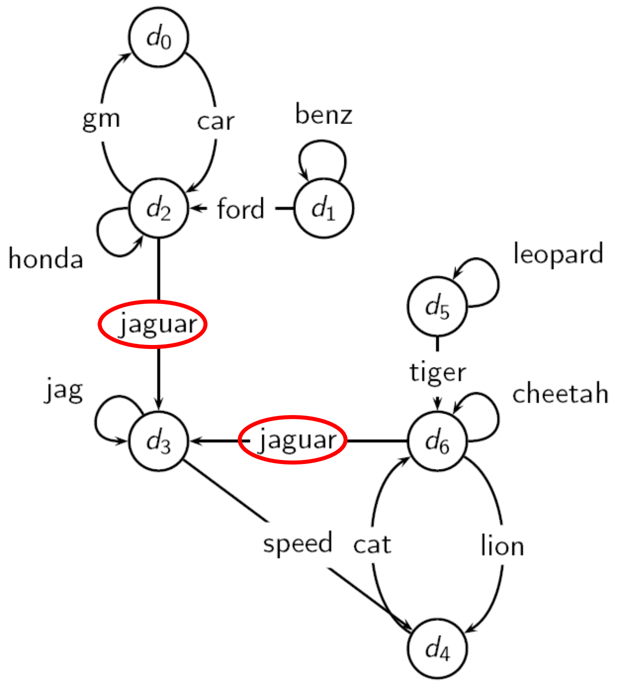
- After convergence:
  - Output pages with the highest h scores as top hubs
  - Output pages with the highest a scores as top authorities
  - So we output two ranked lists

21.3 Hubs and Authorities

- Scaling
  - To prevent the a() and h() values from getting too big, can scale down (归一化) after each iteration
  - Scaling factor doesn't really matter
  - We care about the **relative** (as opposed to absolute) values of the scores

- In most cases, the algorithm converges after a few iterations.

21.3 Hubs and Authorities



Step 0

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	2	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	2	1	0	1

Assuming the query ***jaguar*** and **double-weighting** of links whose anchors contain the query word.

	$\vec{h}_0$	$\vec{h}_1$	$\vec{h}_2$	$\vec{h}_3$	$\vec{h}_4$	$\vec{h}_5$
d_0	0.14	0.06	0.04	0.04	0.03	0.03
d_1	0.14	0.08	0.05	0.04	0.04	0.04
d_2	0.14	0.28	0.32	0.33	0.33	0.33
d_3	0.14	0.14	0.17	0.18	0.18	0.18
d_4	0.14	0.06	0.04	0.04	0.04	0.04
d_5	0.14	0.08	0.05	0.04	0.04	0.04
d_6	0.14	0.30	0.33	0.34	0.35	0.35
	Step 1	Step 3				

(归一化: 和为1)

	$\vec{a}_1$	$\vec{a}_2$	$\vec{a}_3$	$\vec{a}_4$	$\vec{a}_5$	$\vec{a}_6$	$\vec{a}_7$
d_0	0.06	0.09	0.10	0.10	0.10	0.10	0.10
d_1	0.06	0.03	0.01	0.01	0.01	0.01	0.01
d_2	0.19	0.14	0.13	0.12	0.12	0.12	0.12
d_3	0.31	0.43	0.46	0.46	0.46	0.47	0.47
d_4	0.13	0.14	0.16	0.16	0.16	0.16	0.16
d_5	0.06	0.03	0.02	0.01	0.01	0.01	0.01
d_6	0.19	0.14	0.13	0.13	0.13	0.13	0.13
	Step 2 (归一化)	Step 4 (归一化)					

21.3 Hubs and Authorities

- Example

0.85	www.nba.com/bulls
0.25	www.essex1.com/people/jmiller/bulls.htm "da Bulls"
0.20	www.nando.net/SportServer/basketball/nba/chi.html "The Chicago Bulls"
0.15	users.aol.com/rynocub/bulls.htm "The Chicago Bulls Home Page"
0.13	www.geocities.com/Colosseum/6095 "Chicago Bulls"

(Ben-Shaul et al, WWW8)

Authorities for query [Chicago Bulls]

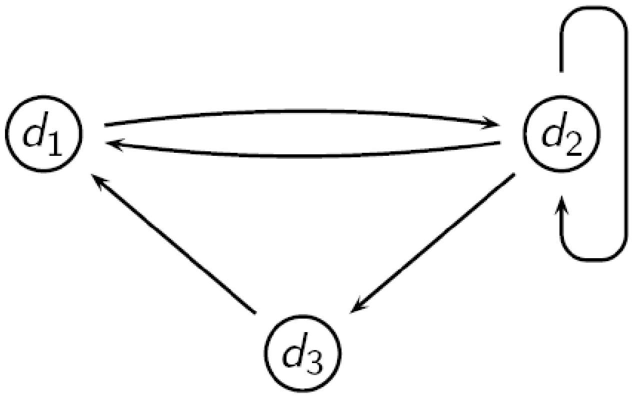
1.62	www.geocities.com/Colosseum/1778 "Unbelieveabulls!!!!!"
1.24	www.webring.org/cgi-bin/webring?ring=chbulls "Erin's Chicago Bulls Page"
0.74	www.geocities.com/Hollywood/Lot/3330/Bulls.html "Chicago Bulls"
0.52	www.nobull.net/web_position/kw-search-15-M2.htm "Excite Search Results: bulls"
0.52	www.halcyon.com/wordsltd/bball/bulls.htm "Chicago Bulls Links"

(Ben-Shaul et al, WWW8)

Hubs for query [Chicago Bulls]

21.3 Hubs and Authorities

- Proof of convergence (1/3)
  - We define an $N \times N$ adjacency matrix A. (We called this the link matrix earlier.)
  - For $1 \leq i, j \leq N$, the matrix entry A_{ij} tells us whether there is a link from page i to page j ($A_{ij} = 1$) or not ($A_{ij} = 0$).
  - Example:



	d_1	d_2	d_3
d_1	0	1	0
d_2	1	1	1
d_3	1	0	0

21.3 Hubs and Authorities

- Proof of convergence (2/3)

  - Define the hub vector $\vec{h} = (h_1, \ldots, h_N)$ as the vector of hub scores. h_i is the hub score of page d_i.
  - Similarly for $\vec{a}$, the vector of authority scores
  - Now we can write $h(d) = \sum_{d \mapsto y} a(y)$ as a matrix operation: $\vec{h} = A\vec{a}$, and we can write $a(d) = \sum_{y \mapsto d} h(y)$ as $\vec{a} = A^T\vec{h}$
  - HITS algorithm in matrix notation:
    - Compute $\vec{h} = A\vec{a}$
    - Compute $\vec{a} = A^T\vec{h}$
    - Iterate until convergence

21.3 Hubs and Authorities

- Proof of convergence (3/3)
  - HITS algorithm in matrix notation. Iterate:
    - Compute $\vec{h} = A\vec{a}$
    - Compute $\vec{a} = A^T\vec{h}$
  - By substitution we get: $\vec{h} = AA^T\vec{h}$ and $\vec{a} = A^TA\vec{a}$
    - Thus, $\vec{h}$ is an eigenvector of AA^T and $\vec{a}$ is an eigenvector of A^TA.
  - So the HITS algorithm is actually a special case of the power method, and hub and authority scores are eigenvector values.
  - HITS and PageRank both formalize link analysis as eigenvector problems.

21.3 Hubs and Authorities

- PageRank can be precomputed
- HITS has to be computed at query time (HITS is too expensive in most application scenarios)

21.3 Hubs and Authorities

- PageRank and HITS make two different design choices concerning
  - (i) the eigen problem formalization
  - (ii) the set of pages to apply the formalization to

  These two are orthogonal (we could also apply HITS to the entire web and PageRank to a small base set)

- Claim: On the web, a good hub is almost always also a good authority.

- The actual difference between PageRank ranking and HITS ranking is therefore not as large as one might expect.

Summary

- 21.1 The Web as a graph
- 21.2 PageRank
- 21.3 Hubs and Authorities
- 21.4 References and further reading