IoT·인공지능·빅데이터 개론 및 실습

빅데이터 응용 – 링크 분석 (2)

서울대학교 컴퓨터공학부
강 유

Contents

Contents

1. Combating Against WebSpam

1. Web Spam: Overview
2. TrustRank: Combating the Web Spam

1 Web Spam: Overview

(1) What is Web Spam?

- Spamming:
  - Any deliberate action to boost a web page's position in search engine results, incommensurate with page's real value
- Spam:
  - Web pages that are the result of spamming
- This is a very broad definition
  - SEO industry might disagree!
  - SEO = search engine optimization
- Approximately 10-15% of web pages are spam

1 Web Spam: Overview

(2) Web Search

- Early search engines:
  - Crawl the Web
  - Index pages by the words they contained
  - Respond to search queries (lists of words) with the pages containing those words
- Early page ranking:
  - Attempt to order pages matching a search query by “importance”
  - First search engines considered:
    - (1) Number of times query words appeared
    - (2) Prominence of word position, e.g. title, header

1 Web Spam: Overview

(3) First Spammers

- As people began to use search engines to find things on the Web, those with commercial interests tried to exploit search engines to bring people to their own site – whether they wanted to be there or not
- Example:
  - Shirt-seller might pretend to be about “movies”
- Techniques for achieving high relevance/importance for a web page

1 Web Spam: Overview

(4) First Spammers: Term Spam

- How do you make your page appear to be about movies?
  - (1) Shirt-seller might pretend to be about “movies” (1) Add the word `movie’ 1,000 times to your page
  - Set text color to the background color, so only search engines would see it
  - (2) Or, run the query “movie” on your target search engine
  - See what page came first in the listings
  - Copy it into your page, make it “invisible”
- These and similar techniques are term spam

1 Web Spam: Overview

(5) Google's Solution to Term Spam

- Believe what people say about you, rather than what you say about yourself
  - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text

- PageRank as a tool to measure the "importance" of Web pages

1 Web Spam: Overview

(6) Why It Works?

- Our hypothetical shirt-seller loses
  - Saying he is about movies doesn't help, because others don't say he is about movies
  - His page isn't very important, so it won't be ranked high for shirts or movies
- Example:
  - Shirt-seller creates 1,000 pages, each links to his page with "movie" in the anchor text
  - These pages have no links in, so they get little PageRank
  - So the shirt-seller can't beat truly important movie pages, like IMDB

1 Web Spam: Overview

(7) Google vs. Spammers: Round 2!

- Once Google became the dominant search engine, spammers began to work out ways to fool Google
- Spam farms were developed to concentrate PageRank on a single page
- Link spam:
  - Creating link structures that boost PageRank of a particular page

1 Web Spam: Overview

(8) Link Spamming

- Three kinds of web pages from a spammer's point of view
  - Inaccessible pages
    - spammer has no control
  - Accessible pages
    - e.g., blog comments pages
    - spammer can post links to his pages
  - Owned pages
    - Completely controlled by spammer
    - May span multiple domain names

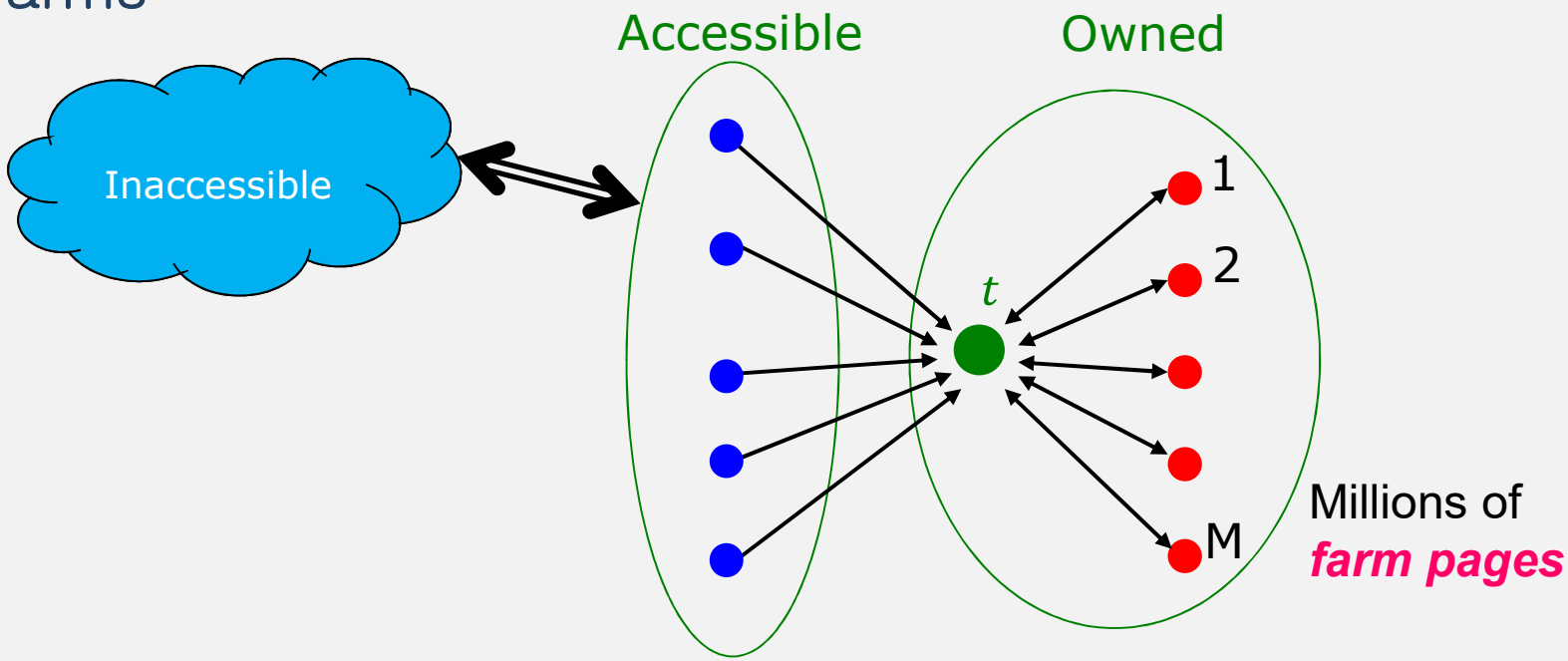
1 Web Spam: Overview

(9) Link Farms

- Spammer's goal:
  - Maximize the PageRank of target page t
- Technique:
  - Get as many links from accessible pages as possible to target page t
  - Construct "link farm" using owned pages to get PageRank multiplier effect

1 Web Spam: Overview

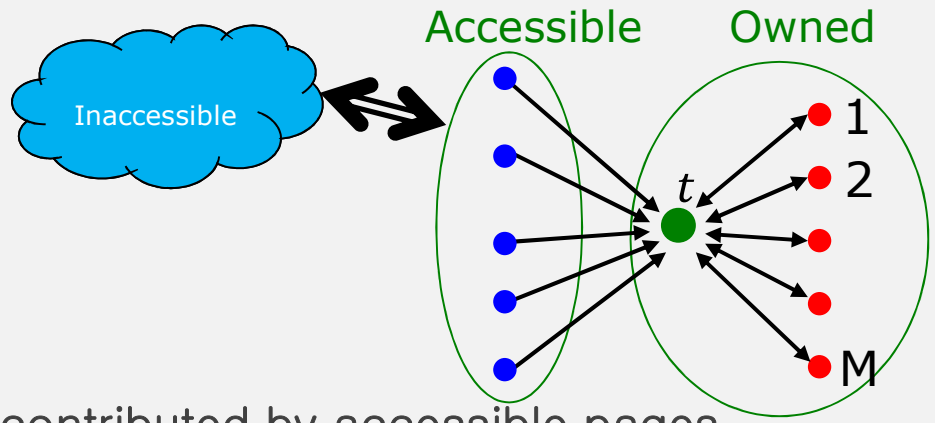
(9) Link Farms



One of the most common and effective organizations for a link farm

1 Web Spam: Overview

(10) Analysis

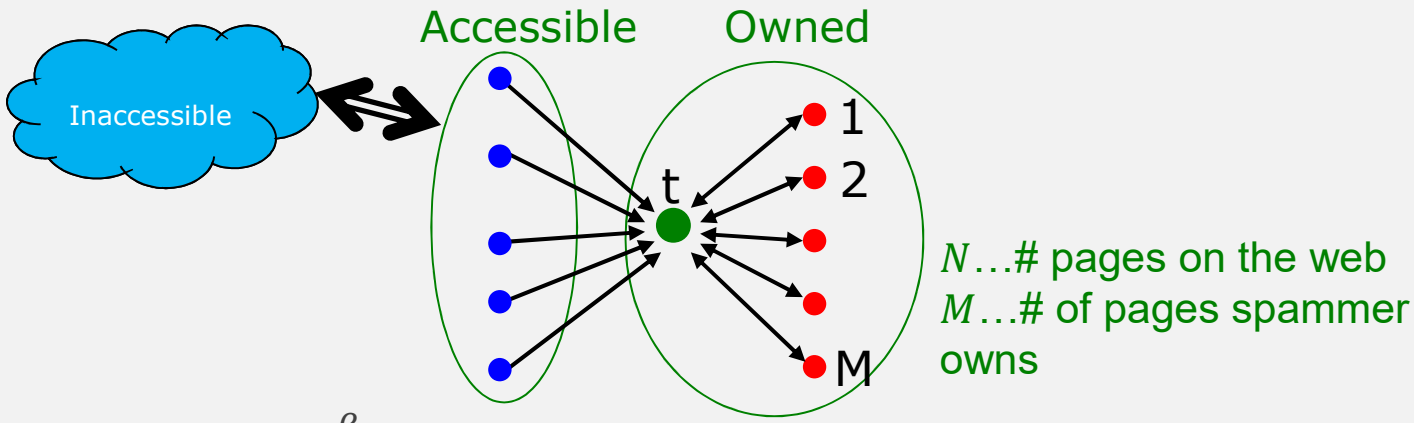


N…# pages on the web
M…# of pages spammer owns

- x: PageRank contributed by accessible pages
- y: PageRank of target page t
- Rank of each "farm" page = $\frac{\beta y}{M} + \frac{1-\beta}{N}$
- $y = x + \beta M \left[\frac{\beta y}{M} + \frac{1-\beta}{N}\right] + \frac{1-\beta}{N} = x + \beta^2 y + \frac{\beta(1-\beta)M}{N} + \frac{1-\beta}{N}$
- $y = \frac{x}{1-\beta^2} + c\frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$

1 Web Spam: Overview

(10) Analysis



N…# pages on the web
M…# of pages spammer owns

- $y = \frac{x}{1-\beta^2} + c\frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$
- For $\beta = 0.85, 1/(1-\beta^2) = 3.6$
- Multiplier effect for acquired PageRank
- By making M large, we can make y as large as we want (up to c)

Contents

② TrustRank: Combating the Web Spam

(1) Combating Spam

Combating term spam

- Use anchor text and PageRank
- Analyze text using statistical methods
- Also useful: Detecting approximate duplicate pages

Combating link spam

- Detection and blacklisting of structures that look like spam farms
  - Leads to another war – hiding and detecting spam farms
- TrustRank = topic-specific PageRank with a teleport set of trusted pages
  - Example: .edu domains, similar domains for non-US schools

② TrustRank: Combating the Web Spam

(2) TrustRank: Idea

- Basic principle: Approximate isolation
  - It is rare for a “good” page to point to a “bad” (spam) page
- Sample a set of seed pages from the web
- Have an oracle (human) to identify the good pages and the spam pages in the seed set
  - Expensive task, so we must make seed set as small as possible

② TrustRank: Combating the Web Spam

(3) Trust Propagation

- Call the subset of seed pages that are identified as good the trusted pages
  - principle: Approximate isolation
- Perform a topic-sensitive PageRank with teleport set = trusted pages
  - Propagate trust through links:
    - Each page gets a trust value between 0 and 1
- Solution 1: Use a threshold value and mark all pages below the trust threshold as spam

② TrustRank: Combating the Web Spam

(4) Why is it a good idea?

Trust attenuation

- The degree of trust conferred by a trusted page decreases with the distance in the graph

Trust splitting:

- The larger the number of out-links from a page, the less trust the page author gives to each out-link
- Trust is split across out-links

② TrustRank: Combating the Web Spam

(5) Picking the Seed Set

- Two conflicting considerations:
  - Human has to inspect each seed page, so seed set must be as small as possible
  - Must ensure every good page gets adequate trust rank, so need to make all good pages reachable from seed set by short paths

❷ TrustRank: Combating the Web Spam

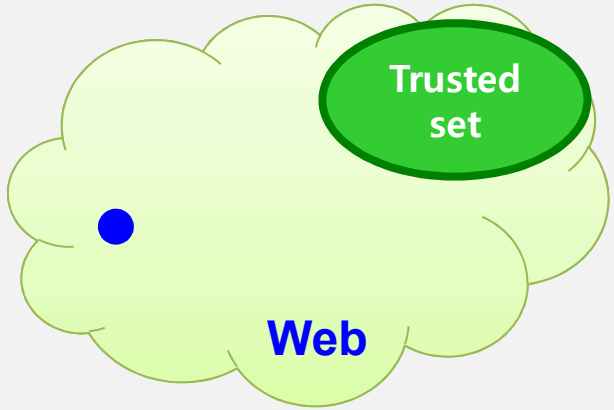
(6) Approaches to Picking Seed Set

- Suppose we want to pick a seed set of k pages
- How to do that?
- (1) PageRank:
  - Pick the top k pages by PageRank
  - Main idea: you can't get a bad page's rank really high
- (2) Use trusted domains whose membership is controlled, like .edu, .mil, .gov

② TrustRank: Combating the Web Spam

(7) Spam Mass

- In the TrustRank model, we start with good pages and propagate trust
- Complementary view:
  - What fraction of a page's PageRank comes from spam pages?
- In practice, we don't know all the spam pages, so we need to estimate



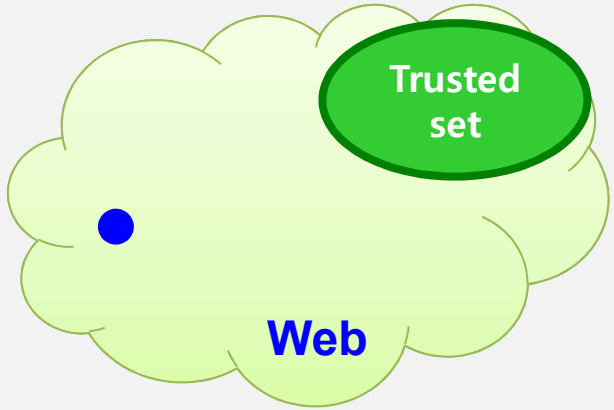
② TrustRank: Combating the Web Spam

(8) Spam Mass Estimation

- Solution
  - $\boldsymbol{r}_p$: PageRank of page p
  - $\boldsymbol{r}_p^+$: PageRank of page p with teleport into trusted pages only
- Then: What fraction of a page's PageRank comes from spam pages?

$$\boldsymbol{r}_p^- = \boldsymbol{r}_p - \boldsymbol{r}_p^+$$

- Spam mass of $p = \frac{\boldsymbol{r}_p^-}{\boldsymbol{r}_p^+}$
  - Pages with high spam mass are spam



Contents

1 HITS: Hubs and Authorities

(1) Hubs and Authorities

- HITS (Hypertext-Induced Topic Selection)
  - Is a measure of importance of pages or documents, similar to PageRank
  - Proposed at around same time as PageRank ('98)
- Goal: Say we want to find good newspapers
  - Don't just find newspapers. Find "experts" – people who link in a coordinated way to good newspapers
- Idea: Links as votes
  - Page is more important if it has more links
    - In-coming links? Out-going links?

1 HITS: Hubs and Authorities

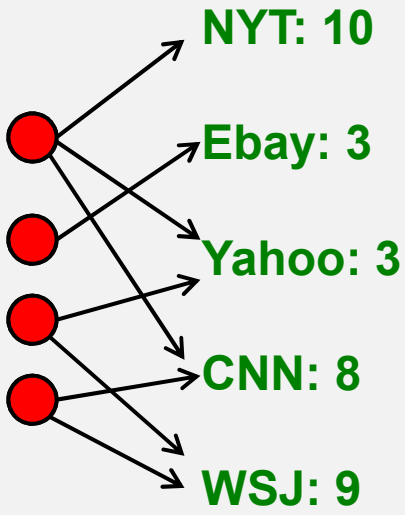
(2) Finding newspapers

- Hubs and Authorities

  Each page has 2 scores:

  - Quality as an expert (hub):
    - Total sum of votes of authorities it points to
  - Quality as a content (authority):
    - Total sum of votes coming from experts

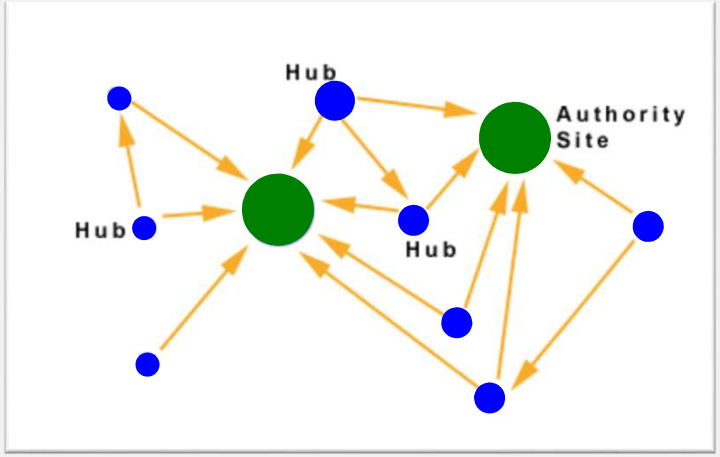
- Principle of repeated improvement



1 HITS: Hubs and Authorities

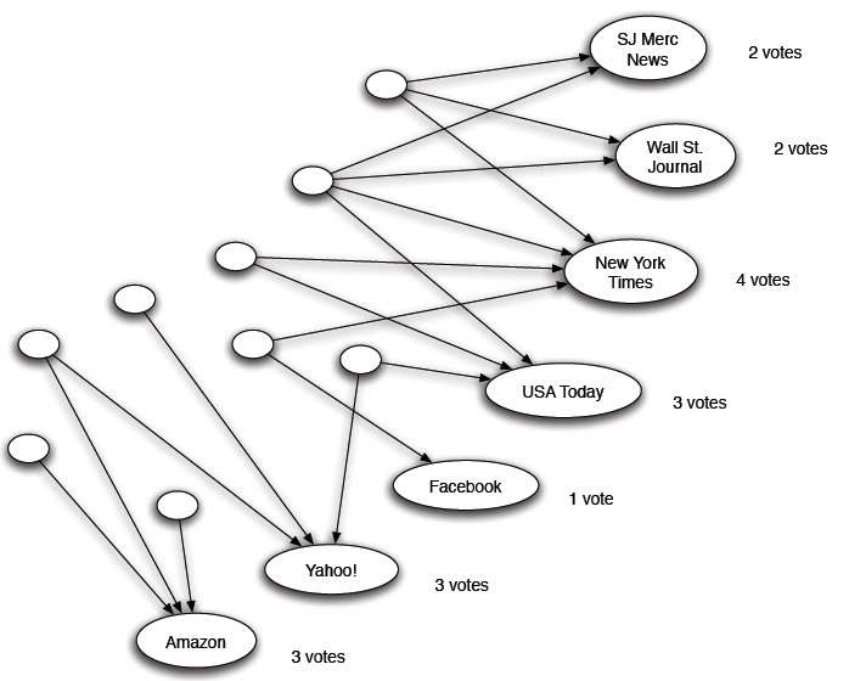
(3) Hubs and Authorities

- Interesting pages fall into two classes:
  - 1) Authorities are pages containing useful information
    - Newspaper home pages
    - Course home pages
    - Home pages of auto manufacturers
  - 2) Hubs are pages that link to authorities
    - List of newspapers
    - Course bulletin
    - List of US auto manufacturers



1 HITS: Hubs and Authorities

(4) Counting in-links: Authority

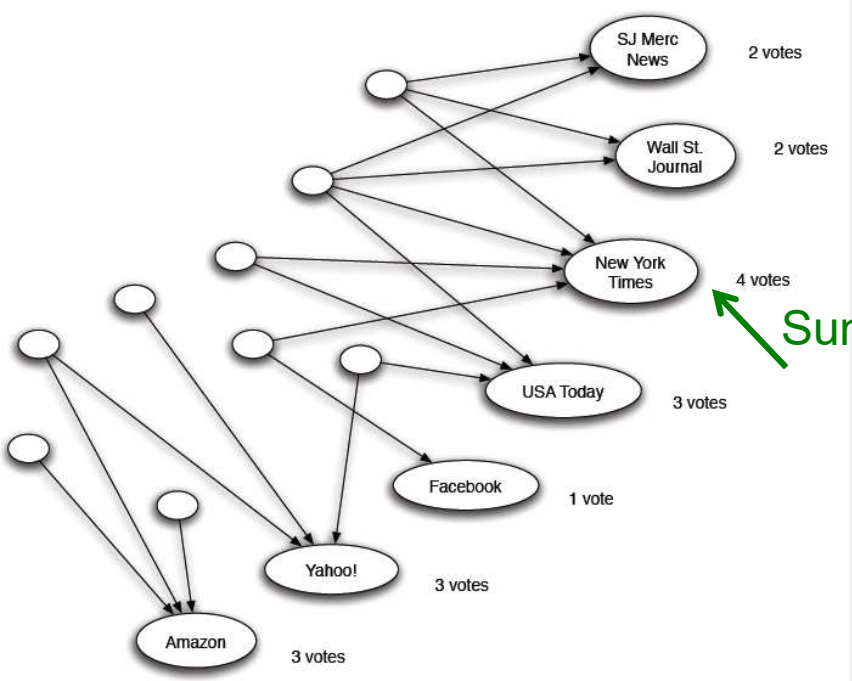


Each page starts with **hub** score 1.
Authorities collect their votes

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

1 HITS: Hubs and Authorities

(4) Counting in-links: Authority



Each page starts with **hub** score 1.
Authorities collect their votes

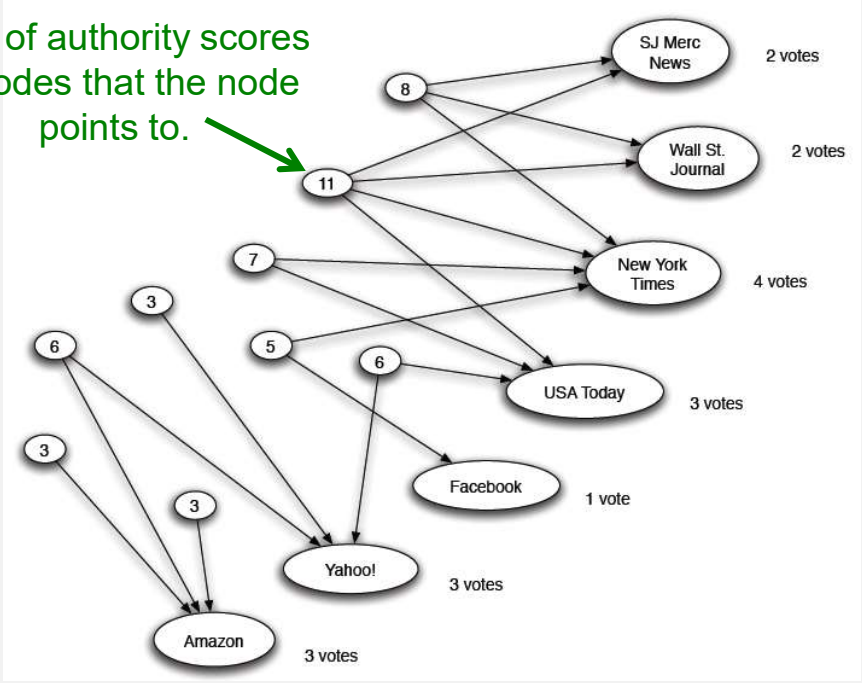
Sum of **hub** scores of nod
es pointing to NYT.

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

1 HITS: Hubs and Authorities

(5) Expert Quality: Hub

Sum of authority scores of nodes that the node points to.

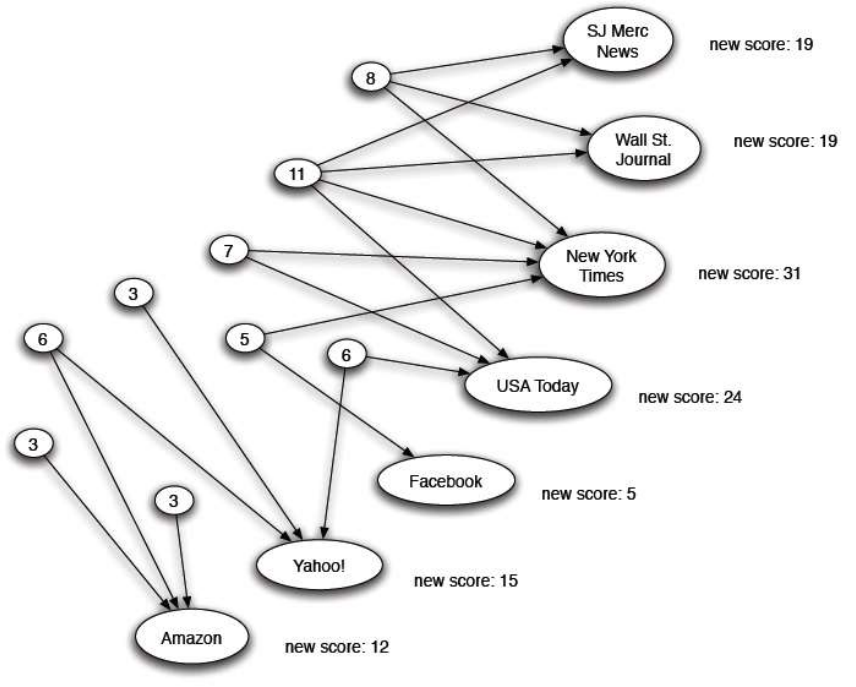


Hubs collect authority scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

1 HITS: Hubs and Authorities

(6) Reweighting



Authorities again collect the **hub** scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

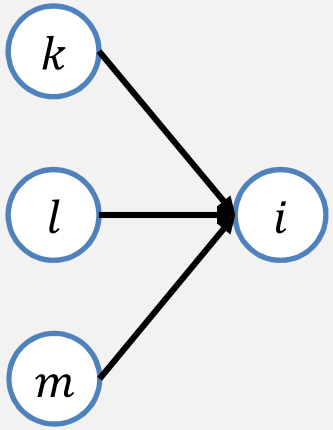
1 HITS: Hubs and Authorities

(7) Mutually Recursive Definition

- A good hub links to many good authorities
- A good authority is linked from many good hubs
- Model using two scores for each node:
  - Hub score and Authority score
  - Represented as vectors $\boldsymbol{h}$ and $\boldsymbol{a}$

1 HITS: Hubs and Authorities

(8) HITS



Then:

$$\boldsymbol{a}_i = \boldsymbol{h}_k + \boldsymbol{h}_l + \boldsymbol{h}_m$$

that is

$\boldsymbol{a}_i$ = sum of $\boldsymbol{h}_j$ over all j that edge (j, i) exis ts

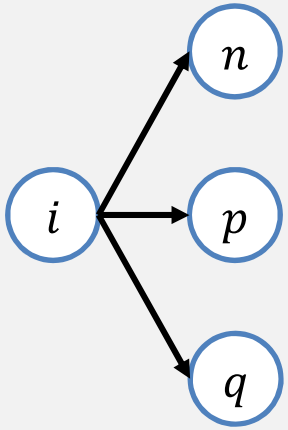
or

$$\boldsymbol{a} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{h}$$

where $\boldsymbol{A}$ is the adjacency matrix (i, j) is 1 if the edge from i to j exists

1 HITS: Hubs and Authorities

(8) HITS



Symmetrically, for the ‘hubness’

$$\boldsymbol{h}_i = \boldsymbol{a}_n + \boldsymbol{a}_p + \boldsymbol{a}_q$$

that is

$\boldsymbol{h}_i$ = sum of $\boldsymbol{a}_j$ over all j that edge (i, j) exis ts

or

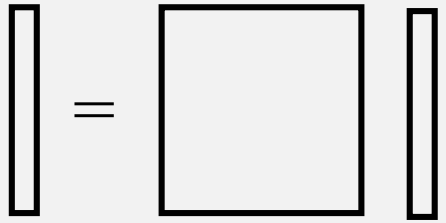
$$\boldsymbol{h} = \boldsymbol{A}\boldsymbol{a}$$

1 HITS: Hubs and Authorities

(8) HITS

- Iterate the following equations until they converge

$$\boldsymbol{h} = \boldsymbol{A}\,\boldsymbol{a}$$
$$\boldsymbol{a} = \boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{h}$$



1 HITS: Hubs and Authorities

(9) Convergence of HITS

- Iterate the following equations until they converge

In short, the solutions to the iterative algorithm

$$\boldsymbol{h} = \boldsymbol{A}\,\boldsymbol{a}$$
$$\boldsymbol{a} = \boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{h}$$

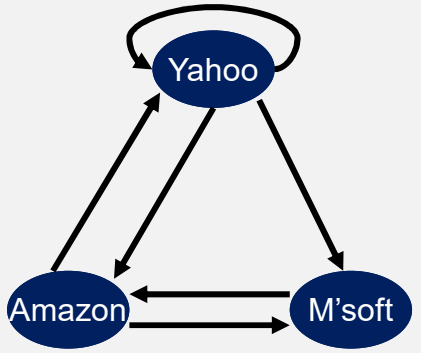
are the largest eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

Starting from random $\boldsymbol{a}'$ and iterating, we'll eventually converge

1 HITS: Hubs and Authorities

(10) Example of HITS

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad A^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$



	iteration	0	1	2	3	⋯	∞
$\boldsymbol{h}$(yahoo)		0.58	0.80	0.80	0.79	⋯	0.788
$\boldsymbol{h}$(amazon)	=	0.58	0.53	0.53	0.57	⋯	0.577
$\boldsymbol{h}$(m′soft)		0.58	0.27	0.27	0.23	⋯	0.211

	iteration	0	1	2	3	⋯	∞
$\boldsymbol{a}$(yahoo)		0.58	0.58	0.62	0.62	⋯	0.628
$\boldsymbol{a}$(amazon)	=	0.58	0.58	0.49	0.49	⋯	0.459
$\boldsymbol{a}$(m′soft)		0.58	0.58	0.62	0.62	⋯	0.628

Note that each vector is scaled so that L2 norm is 1

SUMMARY

Motivation for link analysis

- Graphs are everywhere
- Web as graphs

Pagerank: an important graph ranking algorithm

- A page is important if it is pointed to by other important pages
- Pagerank vector gives the stationary distribution for the random walk on a graph

Topic-specific Pagerank

- Teleports to a topic specific set of pages
- Useful when query nodes are given

SUMMARY

Combating against WebSpam

- WebSpam: definition and method of attacks
  - Term Spam and Link Spam
- TrustRank: how to combat WebSpam
  - Topic-sensitive PageRank with teleport set = trusted pages

HITS algorithm: another algorithm to rank pages, but with two scores (authority and hub scores)

- Authorities are pages containing useful information
- Hubs are pages that link to authorities

Questions?