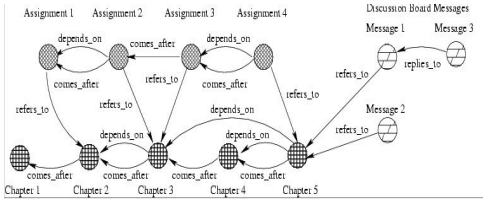
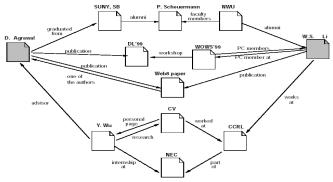
Web ….

Web ….



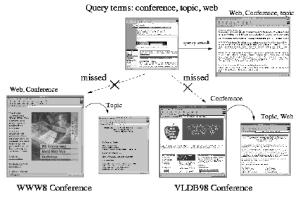
Web ….

- Approach 1: use standard IR techniques to find pages that satisfy a query



Web ….

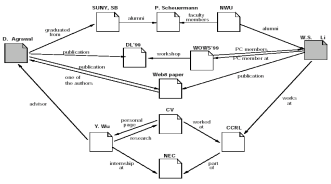
- Approach 1: use standard IR techniques to find pages that satisfy a query



isa Sapino (BDM 2018)

Web ….

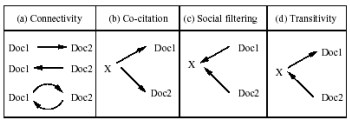
- Approach 2: integrate IR techniques with structure/link analysis



Maria Luisa Sapino (BDM 2018)

Web ….

- Approach 2: integrate IR techniques with structure/link analysis

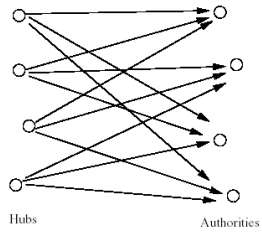


Maria Luisa Sapino (BDM 2018)

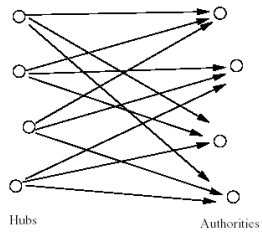
HITS algorithm

- Good pages are categorized into two types
  - Hubs: point to many pages of high quality
  - Authorities: pages of high quality

Hubs and authorities



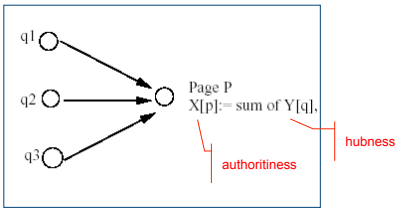
Hubs and authorities



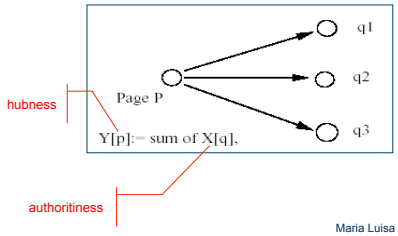
•Good hubs should point to good authorities

•Good authorities must be pointed by good hubs.

Topic distillation by iterative mutual reinforcement



Topic distillation by iterative mutual reinforcement



HITS

- Use IR to find the candidate pages

HITS

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set

HITS

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set
- Compute authority and hub values for all pages (iterate!!)

$$a(i) = \sum_{j \in in(i)} h(j) \qquad h(i) = \sum_{j \in out(i)} a(j)$$

HITS

- Matrix notation

$$\vec{a} = E^T \vec{h} \qquad \vec{h} = E \vec{a}$$

…reminder

- Eigenvalue and eigenvector
- Given a matrix E, let *c* (scalar) and *x* (vector) be such that

$$c\vec{x} = E\vec{x}$$

Eigenvalue

Eigenvector

…authorities

$$\vec{a} = E^T \vec{h}$$

…authorities

$$\vec{a} = E^T E \vec{a}$$

a is an eigenvector of $E^T E$

…hubs

$$\vec{h} = EE^T \vec{h}$$

h is an eigenvector of EE^T

HITS and LSI???

- …reminder: SVD of *E* is

$$E = UMV^T$$

where

- *M* is a diagonal matrix
- $U^TU=I$
- $V^TV=I$

..then

$$EE^T = UMV^TVMU^T$$

..then

$$EE^T = UMV^T VMU^T$$

$$EE^T = UMIMU^T$$

$$EE^T = UM^2U^T$$

…then again

$$EE^T = UM^2U^T$$

$$EE^TU = UM^2U^TU = UM^2$$

…then again

$$EE^T = UM^2U^T$$

$$EE^TU = UM^2U^TU = UM^2$$

$$EE^TU_j = U_j m_j^{\,2}$$

Eigenvector

Eigenvalue

…in fact

$$EE^T U_j = U_j m_j^2$$

Eigenvector — Eigenvalue

$$EE^T \vec{h} = \vec{h}$$

$h = U_1$

…in fact

$$EE^T U_j = U_j m_j^2$$

Eigenvector — Eigenvalue

$$EE^T \vec{h} = \vec{h}$$

$h = U_1$

HITS is similar to LSI, but on (source, destination) rather than (term,document) matrix

Problem

- Topic drift
  - Pages include neighbors
  - Neighbors may be good hubs, authorities; but may not have good content match

Clever system

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set
- Compute authority and hub values for all pages (iterate!!)
- Consider text next to the link!!!!

$$a(i) = \sum_{j \in in(i)} w(j \to i) h(j) \qquad h(i) = \sum_{j \in out(i)} w(i \to j) a(j)$$

Problem

- Topic drift
  - Pages include neighbors
  - Neighbors may be good hubs, authorities; but may not have good content match
- Slow
  - Iteration is not good!
  - One eigenvector computation per query

PageRank

- Random Surfer
  - Jumps from page to page with uniform probability
  - Occasionally jump to a random page with small probability (1-β)
  - If no out page, then jump to any page with equal probability

PageRank

- Random Surfer (N pages)
  - Jumps from page to page with uniform probability
  - Occasionally jump to a random page with small probability (1-β)
  - If no out page, then jump to any page with equal probability

$$\mathbf{Z} = (1-\beta)\left[\frac{1}{N}\right]_{N\times N} + \beta\mathbf{M}$$

Transition matrix

$$M_{ji} = \begin{cases} \frac{1}{|out(i)|} & \text{if there is an edge from i to j} \\ 0 & \text{otherwise} \end{cases}$$

PageRank

- Random Surfer (N pages)
  - Jumps from page to page with uniform probability
  - Occasionally jump to a random page with small probability (1-β)
  - If no out page, then jump to any page with equal probability

$$P(j) = \frac{1-\beta}{N} + \beta \sum_{i \in in(j)} \frac{P(i)}{out(i)}$$

Probability that the surfer is at page j

PageRank

- Random Surfer (N pages)
  - Jumps from page to page with uniform probability
  - Occasionally jump to a random page with small probability (1-β)
  - If no out page, then jump to any page with equal probability

$$P(j) = \frac{1-\beta}{N} + \beta \sum_{i \in in(j)} \frac{P(i)}{out(i)}$$

Probability that the surfer is at page j

Primary eigenvector of the transition matrix **Z**