UNIVERSITY of DELAWARE

Link Analysis

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

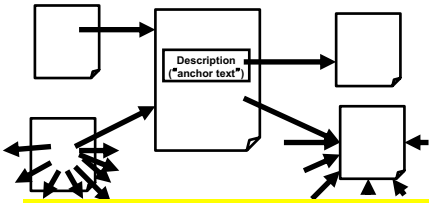
UNIVERSITY of DELAWARE

Exploiting Inter-document Structures

- Web pages have links.
- Challenge: how to exploit links to improve ranking?
  - Anchor texts
  - PageRank
  - HITS

UNIVERSITY of DELAWARE

Web pages form a directed graph.



Assumption : The anchor of the hyperlink describes the target page (textual context)

UNIVERSITY of DELAWARE

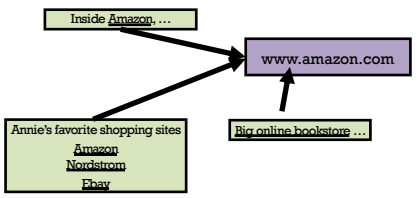
Why is Anchor Text Useful?

- For query *"amazon"*, how to distinguish between:
  - Amazon's homepage
  - Amazon's policy page
  - a spam page

UNIVERSITY of DELAWARE

Indexing anchor text

- When indexing a document *D*, include anchor text from links pointing to *D*.



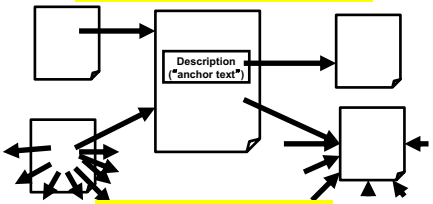
UNIVERSITY of DELAWARE

Indexing anchor text

- It can sometimes have unexpected side effects
  - Google Bomb
  - *e.g., miserable failure.*
- The anchor text can be weighted based on the authority of the anchor page.
  - Trust webpages from cnn.com than those from 123.com

Web pages form a directed graph.

Basic idea: Peer endorsement



Links indicate the utility of a doc

Using link structure to measure page importance

- Simplest solution
  - use link counts as popularity measure
    - Page score = number of in-links.

Limitation: can be easily spammed to give a page a high score

PageRank: Capturing Page "Popularity"

- PageRank improves over simple counting
  - Consider indirect endorsement
    - If a page is endorsed by a highly endorsed webpage, it counts more than being endorsed by a webpage without any endorsements.

A simple PageRank scoring

- Random work on web pages:
  - Start at a random page
  - At each step, following one of the links in the current page and move on to the next page

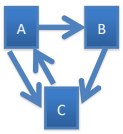
- Eventually, each page has a long-term visit rate, which is used as the page's score.

An Example of simple PageRank

- $PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$
- More generally,

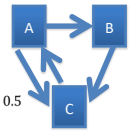
$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  - where B_u is the set of pages that point to u, and L_v is the number of outgoing links from page v



Simple PageRank – Initialization

- Initialize them with the equal values
  $PR(A) = PR(B) = PR(C) = \frac{1}{3}$
- First iteration
  $PR(A) = 0.33, PR(B) = 0.17, PR(C) = 0.17 + 0.33 = 0.5$
- Second iteration
  $PR(A) = 0.5, PR(B) = 0.17, PR(C) = 0.33/2 + 0.17 = 0.33$
- Third iteration
  $PR(A) = 0.5, PR(B) = 0.17, PR(C) = 0.33/2 + 0.17 = 0.33$
- Until Converge
  - PR(C) = 0.4, PR(A) = 0.4, and PR(B) = 0.2

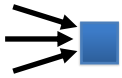


Any limitations with this method?

UNIVERSITY of DELAWARE UNIVERSITY of DE

Limitation: Dangling Links

- Random walk can stuck at the dead ends.
- Eventually, the score of the dead end would be 1, and the scores of other webpages would be 0.



UNIVERSITY of DELAWARE UNIVERSITY of DE

Solution: Teleporting

- In addition to following the links, we can jump to a random web page.
- Given a web page,
  - with probability 10%, jump to a random web page.
  - With remaining probability (90%), follow one of the outlinks.

16

UNIVERSITY of DELAWARE UNIVERSITY of DE

PageRank: Capturing Page "Popularity"

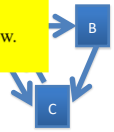
- PageRank improves over simple counting
  - Consider indirect citations
  - Being referred by a highly referred web page counts more…
- PageRank can also be interpreted as random surfing.

17

UNIVERSITY of DELAWARE UNIVERSITY of DE

PageRank

Random surfing model:
At any page,
With probability λ, randomly jumping to a page
With probability (1- λ), randomly picking an outlink to follow.
PageRank of d= average probability of visiting page d



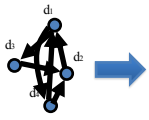
- More generally,

$$PR(u) = \frac{\lambda}{N} + (1-\lambda).\sum_{v \in B_u} \frac{PR(v)}{L_v}$$

– where N is the number of pages, λ is a parameter.

UNIVERSITY of DELAWARE UNIVERSITY of DE

PageRank – Matrix Representation



"Transition matrix"

$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

Mij =probability of going from di to dj

$$\sum_{j=1}^{N} M_{ij} = 1$$

Probability of visiting page di at time t

$$p_{t+1}(d_j) = (1-\alpha)\sum_{i=1}^{N} M_{ij} p_t(d_i) + \alpha \sum_{i=1}^{N} \frac{1}{N} p_t(d_i)$$

Probability of visiting page dj at time t+1

Reaching dj via following a link

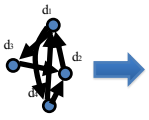
Reaching dj via random jump

N= # of pages (nodes)

20

UNIVERSITY of DELAWARE UNIVERSITY of DE

PageRank – Matrix Representation (Cont.)



"Transition matrix"

$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

Mij =probability of going from di to dj

$$\sum_{j=1}^{N} M_{ij} = 1$$

$$p_{t+1}(d_j) = (1-\alpha)\sum_{i=1}^{N} M_{ij} p_t(d_i) + \alpha \sum_{i=1}^{N} \frac{1}{N} p_t(d_i)$$

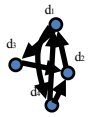
$$p(d_j) = \sum_{i=1}^{N} [\frac{1}{N}\alpha + (1-\alpha)M_{ij}]p(d_i)$$

$$\bar{p} = (\alpha I + (1-\alpha)M)^T \bar{p} \qquad I_{ij} = \frac{1}{N}$$

This can be solved with an iterative algorithm.

21

PageRank – An Example



$$p(d_j) = \sum_{i=1}^{N} [\frac{1}{N}\alpha + (1-\alpha)M_{ij}]p(d_i)$$

$$\bar{p} = (\alpha I + (1-\alpha)M)^T \bar{p} \qquad I_{ij} = \frac{1}{N}$$

$$A = (1-0.2)M + 0.2I = 0.8\begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} + 0.2\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

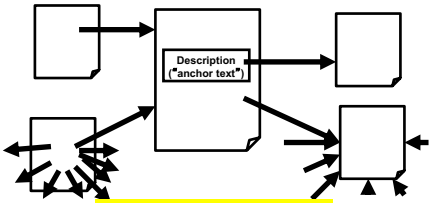
$$\begin{bmatrix} p_{t+1}(d_1) \\ p_{t+1}(d_2) \\ p_{t+1}(d_3) \\ p_{t+1}(d_4) \end{bmatrix} = A^{\mathrm{T}} \begin{bmatrix} p_t(d_1) \\ p_t(d_2) \\ p_t(d_3) \\ p_t(d_4) \end{bmatrix} = \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.45 \\ 0.05 & 0.05 & 0.85 & 0.45 \\ 0.45 & 0.05 & 0.05 & 0.05 \\ 0.45 & 0.05 & 0.05 & 0.05 \end{bmatrix} \begin{bmatrix} p_t(d_1) \\ p_t(d_2) \\ p_t(d_3) \\ p_t(d_4) \end{bmatrix}.$$

HITS: Capturing Authorities & Hubs

- Intuitions
  - Pages that are widely cited are good authorities
  - Pages that cite many other pages are good hubs
- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - Iterative reinforcement…

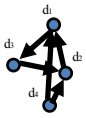
Web pages form a directed graph.

Basic idea: Peer endorsement



Hub **Links indicate the utility of a doc** Authority

The HITS Algorithm



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

"Adjacency matrix"

Initial values: $a(d_i)=h(d_i)=1$

$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$
$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

} Iterate

$$\bar{h} = A\bar{a}; \quad \bar{a} = A^T\bar{h}$$
$$\bar{h} = AA^T\bar{h}; \quad \bar{a} = A^TA\bar{a}$$

Normalize:

$$\sum_i a(d_i)^2 = \sum_i h(d_i)^2 = 1$$