Part I: Web Structure Mining
Chapter 2: Hyperlink Based Ranking

- Social Network Analysis
- PageRank
- Authorities and Hubs
- Link Based Similarity Search
- Enhanced Techniques for Page Ranking

Social Networks

- Directed graph with weights assigned to the edges
- Nodes represent documents and edges – citations from one document to other documents.
- *Prestige* can be associated with the number of input edges to a node (in-degree).
- Prestige has a *recursive* nature – it depends on the authority (or again, the prestige) of citations.

Prestige Score

- Adjacency matrix A

  $A(u,v) = 1$ if document *u* cites document *v*

  $A(u,v) = 0$ otherwise

- *Prestige score*

$$p(u) = \sum_v A(v,u)\, p(v)$$

Computing Prestige Score

- Solving matrix equation

$$P' = A^T P$$

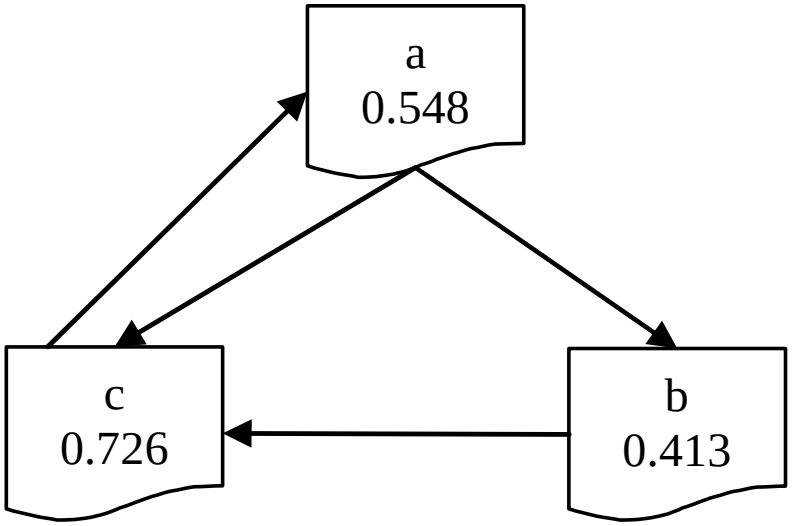
- *Eigen decomposition*

$$\lambda P = A^T P$$

Eigenvector P

Eigenvalue λ

Social Networks Example



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

$$\lambda P = A^T P$$

$$\lambda = 1.325$$

$$P = (0.548 \quad 0.414 \quad 0.726)^T$$

$$1.325 \begin{pmatrix} 0.548 \\ 0.414 \\ 0.726 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.548 \\ 0.414 \\ 0.726 \end{pmatrix}$$

Computing Prestige by Power Iteration

- $P \leftarrow P_0$
- *Loop:*

$$Q \leftarrow P$$

$$P \leftarrow A^T Q$$

$$P \leftarrow \frac{1}{\|P\|} P$$

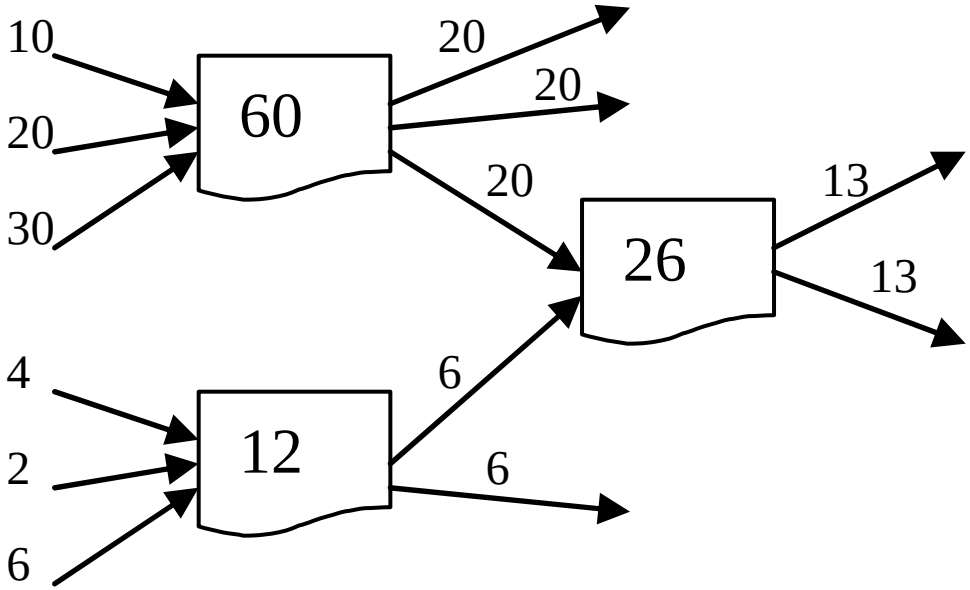
- *While* $\|P - Q\| > \varepsilon$

PageRank

- "Random web surfer" keeps clicking on hyperlinks at random with uniform probability
- Implements *random walk* on the web graph
- If page *u* links to N_u web pages and *v* is one of them then:
  - Once the surfer is at page *u* the probability of visiting page *v* will be $1/N_u$
  - The amount of prestige that page *v* receives from page *u* is $1/N_u$ of the prestige of *u*

Page Rank Propagation

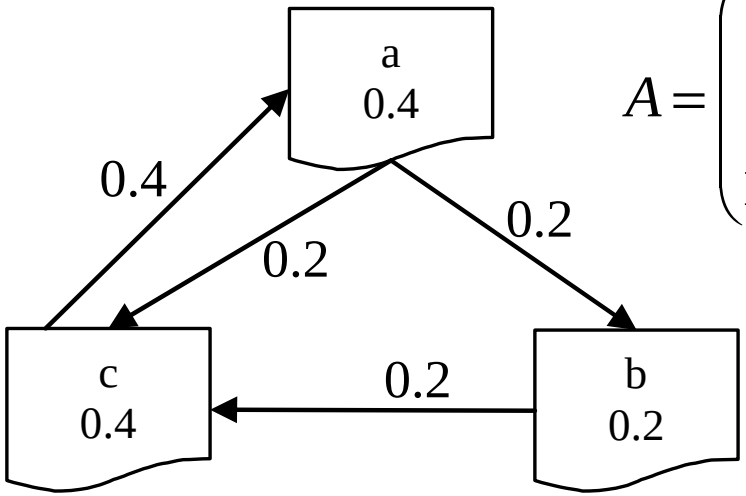
Propagation of page rank $R(u)$



$$R(u) = \lambda \sum_{v} \frac{A(v,u)R(v)}{N_v}$$

$$N_v = \sum_{w} A(v,w)$$

Calculation of PageRank



$$A = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\lambda P = A^T P$$

$$\begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix}$$

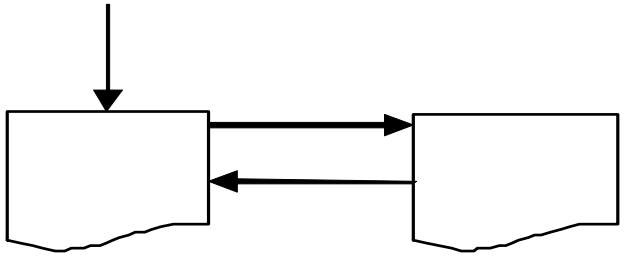
with norm $\|X\|_1 = x_1 + x_2 + \ldots + x_n$

$P^T = (0.666 \;\; 0.333 \;\; 0.666)$ with norm $\|X\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

$P^T = (2 \;\; 1 \;\; 2)$ in integers

Rank Sink and Power Iteration

Rank Sink



Rank source *E(u)*

$$R(u) = \lambda\left(\sum_{v} \frac{A(v,u)R(v)}{N_v} + E(u)\right)$$

$$R \leftarrow R_0$$

$$Q \leftarrow R$$

Loop:

$$R \leftarrow A^T Q$$

$$d \leftarrow \|Q\|_1 - \|R\|_1$$

$$R \leftarrow R + d\,E$$

While $\|R - Q\|_1 > \varepsilon$

PageRank Discussion

- The rank vector *R* defines the probability distribution of a random walk on the Web graph.
- With some low probability the surfer jumps to a random page chosen according to distribution *E*.
- *E* is usually chosen as a uniform vector with a small norm.
- If the norm of *E* is larger the surfer jumps to a random page more often.
- A larger norm of *E* means less contribution of the link structure to the final PageRank score (the distribution of *R* gets closer to *E*).
- The rank source E allows PageRank to be adjusted for *customized ranking* or to avoid *commercial manipulation*.
- Other PageRank applications include *estimating Web traffic*, *optimal crawling* and *web page navigation*.

Authorities and Hubs

- There are problems with using only the in-degree based authority (e.g. some links have noting to do with authority).
- Neither content-based relevance nor link-based authority can do the job alone, rather a good balance between the two is needed.
- Hyperlink Induced Topic Search (HITS) combines content-based relevance with link-based authority ranking.
- Focuses on relevant pages first and then computes authority.
- Works with much smaller and query dependent part of the Web graph.
- Takes into account *hub pages* (pages that point to multiple relevant authoritative pages).

Hyperlink Induced Topic Search (HITS)

- Given a query q a standard IR system finds a small set of relevant web pages called a *root set* R_q.
- The root set is expanded to a *base set* S_q by adding pages that point to and are pointed to by pages from the root set.
- The hyperlink structure of the base set is analyzed to find *authorities* and *hubs*.

Finding Authorities and Hubs

$E(u,v)$ – adjacency matrix of the base set S_q

$X = (x_1\, x_2 \ldots x_n)$ – authority vector

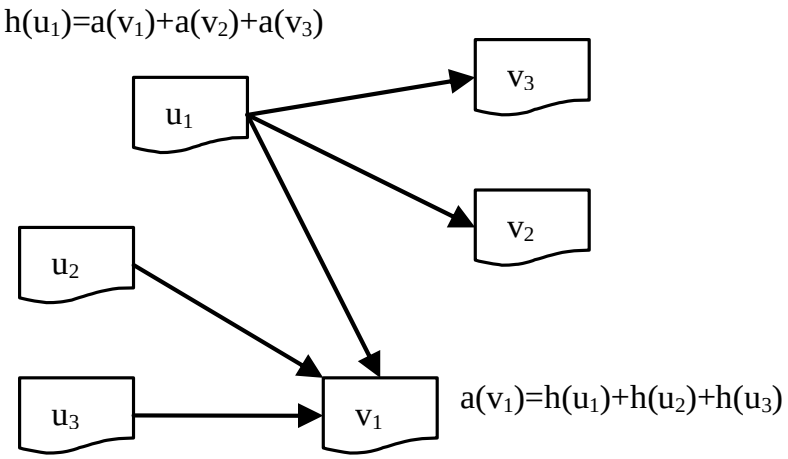
$Y = (y_1\, y_2 \ldots y_n)$ – hub vector

k – tuned parameter

- $X \leftarrow (11\ldots1)$
- $Y \leftarrow (11\ldots1)$
- Loop k times
  - $x_u \leftarrow \sum_{\{v,E(v,u)=1\}} y_v$, for $u = 1, 2, \ldots, n$
  - $y_u \leftarrow \sum_{\{v,E(u,v)=1\}} x_v$, for $u = 1, 2, \ldots, n$
  - normalize X and Y by the L_2 norm
- End loop

Authority score $x_i = a(u_i)$

Hub score $y_i = h(u_i)$



Link-Based Similarity Search

- Find *k* pages pointing to page *u* and use them to form the root set R_u
- Using R_u find the base set S_u
- Compute authorities and hubs in S_u
- Report the highest ranking authorities and hubs as similar pages to *u*.

Enhanced Page Ranking

- *Topic Generalization* (expansion of a set of pages by a number of links)
  - Expansion by one link is used by HITS
  - Expansion by more than one link usually leads to *topic drift*
- *Nepotistic links* (densely linked pages located on a single site or related sites)
  - Assign weights of to inlinks from pages belonging to a single site
- *Outliers* (relevant pages retrieved by keyword search, but far from the central topic of the query)
- Eliminating outliers by *clustering*
  - Create vector space representation for the pages from the root set.
  - Find the *centroid* of the root set (the page that minimizes its cosine similarity to all pages in the set)
  - When expanding the root set discard pages that are too far from the centroid page.