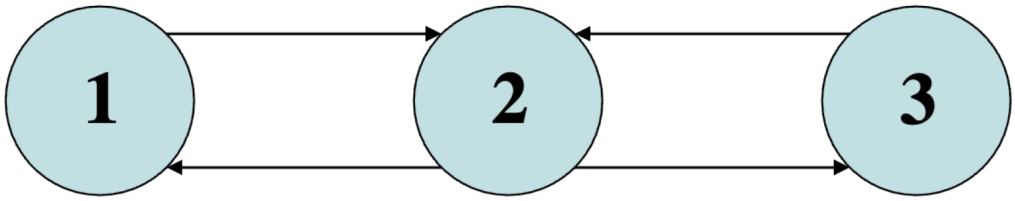
Web

Seed pages -> URLs crawled and parsed <- URLs frontier <-> Unseen Web
Relevance (tf-idf) X Importance (authority)
Front queue: n queues for different priorities (based on quality and frequency of updated)
Back queue: a queue per host

Link Analysis

Start at a random page
At each step, with probability $1 - \alpha$, go out of the current page along one of the links on that page, equi-probably (randomly click on an out-going link)
At each step, with probability α, jump to a random page (input a URL to the browser) "teleporting"
At a dead end, jump to a random web page
$P_{i,j}$ - if at i, probability of going to j
N - № of pages
$P_{i,j} = \frac{1}{N}$ - if dead end
L_i - № of outgoing links
$P_{i,j} = \frac{a}{N} + \frac{1-a}{L_i}$ - if there is link from i to j
$P_{i,j} = \frac{a}{N}$ - else

- Represent the random walk as a Markov chain with teleporting probability $\alpha=0.5$
- Note that teleporting destinations from a webpage includes itself.



$$\begin{bmatrix} 1/6 & 1/6+1/2 & 1/6 \\ 1/6+1/4 & 1/6 & 1/6+1/4 \\ 1/6 & 1/6+1/2 & 1/6 \end{bmatrix}$$

$P_{i,j} \neq 0$ - network is ergodic (can directly go from any to any)

$x = [x_1, \ldots x_n] = 1$ - walk is in i with probability x_i (probability row vector)

Exercise

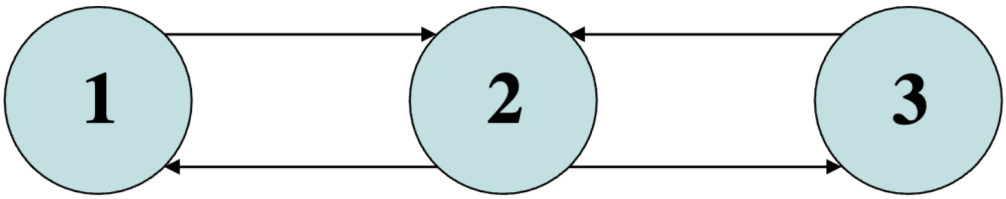
– We start from state 1:

$x = [1, 0, 0]$

– Next step?

$x = [1/6, 2/3, 1/6]$

This means: there's 1/6 chance that the random surfer is at state 1…



$$\begin{bmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{bmatrix}$$

$xP = x$ - steady state

Hub pages are good lists of links on a subject.

Authority pages occur recurrently on good hubs for the subject.

Get all pages containing the query (root set)

Add any page that points/is pointed to root set (base set)

Eliminate links between two pages of same host

h(1),....-> output highest hubs and highest a()

Given text query (say browser), use a text index to get all pages containing browser.

- Call this the root set of pages.

Add in any page that either

- points to a page in the root set, or
- is pointed to by a page in the root set.
- Call this the base set.

To eliminate purely navigational links:

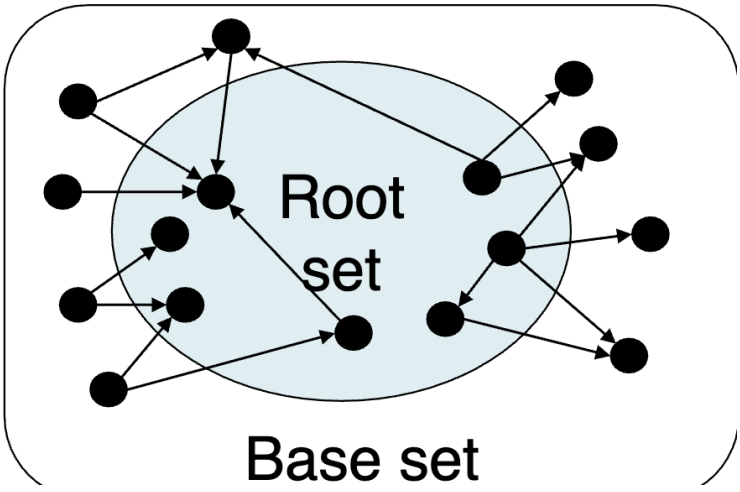
- Eliminate links between two pages on the same host.

Root set

- 200-1000 nodes.

Base set

- up to 5000 nodes.



Initialize: for all x, $h(x) \leftarrow 1$; $a(x) \leftarrow 1$;

Iteratively update all $h(x)$, $a(x)$;

After iterations

- output pages with highest $h()$ scores as top hubs
- highest $a()$ scores as top authorities.

$$h(x) = \sum_{x \to y} a(y)$$

$$a(x) = \sum_{x \to y} h(y)$$

Classification

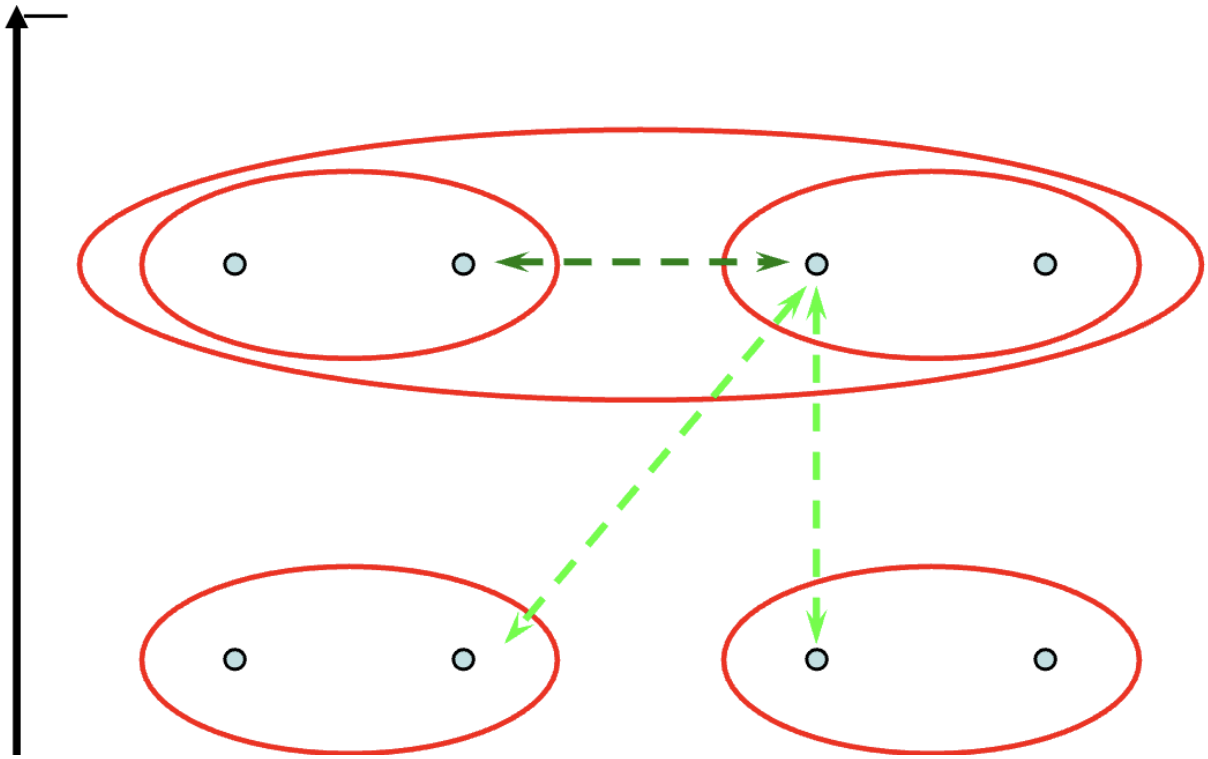
Manual, Automatic (rule-based), Supervised learning (kNN)

Clustering

Unsupervised
Flat algorithms (random->refine) - k-means
Hierarchical (top-down or bottom-up)
Single-link - Similarity of the most cosine-similar
Complete-link - Similarity of the “furthest” points, the least cosine-similar
Centroid - Clusters whose centroids (centers of gravity) are the most cosine-similar
Average-link - Average cosine between pairs of elements
$\frac{|largest\ group|}{|total|}$ - Purity of class

Single Link

– Similarity of the “nearest” points



Social Network Analysis

Information hub
Graph/Adjacency matrix/Adjacency list
Geodesic distance - № of nodes in shorter path between A&B (i.e, 1->2->3 = 2)
In-degree, out-degree - sum of connections to/out from node
$\frac{deg(v)}{n-1}$ - centrality (average distance to all/presence in geodesics)
Reciprocity - fraction of directed edges with inverse
Popularity is used instead of importance
Collaborative filtering - personal tastes are correlated
$\frac{|edges|}{|possible\ edges|=\frac{|nodes|(|nodes|-1)}{2}}$ - density
$\frac{2*|triangles\_of\_neighbors|}{|degree|*(|degree|-1)}$ - Clustering Coefficient

TODO

Review exam prep and quiz 5, quiz 6
Be careful when calculating HITS, PageRank and Clustering Coefficient
Create short notes and Print them