Lecture 9:
I: Web Retrieval
II: Webology

Johan Bollen
Old Dominion University
Department of Computer Science

jbollen@cs.odu.edu
http://www.cs.odu.edu/~jbollen
WWW retrieval

Two approaches

1. Document collection
   (a) Collect documents
   (b) Match terms

2. Citation graphs and social network
   (a) Collect documents
   (b) Apply graph analysis
   (c) Determine page rank by “prestige” indexes
Building a WWW collection: Crawling

1. Distributed collection
   (a) Lack of centralized archive
      i. No central list of URLs
      ii. No or little metadata
   (b) Lack of homogeneous markup
      i. Multimedia

2. Crawling
   (a) Application “walks” network
   (b) Collect documents
   (c) Parse hyperlinks
   (d) Traverse hyperlinks: collect new pages
Crawling Issues

<table>
<thead>
<tr>
<th>Scale</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rapid expansion of collection</td>
<td>1. WWW is still growing</td>
</tr>
<tr>
<td>2. Network delays</td>
<td>(a) Adequate sample?</td>
</tr>
<tr>
<td>(a) Resolving hostname</td>
<td>(b) Scalability issues</td>
</tr>
<tr>
<td>(b) Server request</td>
<td>2. Spider traps</td>
</tr>
<tr>
<td>(c) Checking URL cache</td>
<td>3. Duplication</td>
</tr>
<tr>
<td>3. Parsing and storing page</td>
<td>4. Freshness</td>
</tr>
</tbody>
</table>
Basic Crawler
Perl LWP module

dean - johan/src
Power Crawler

FIGURE 2.2 Typical anatomy of a large-scale crawler.
Distributed Crawler Architectures

**Figure 2.4** Large-scale crawlers often use multiple ISPs and a bank of local storage servers to store the pages crawled.
Focused crawling

1. Subject matters
   (a) Normal crawlers:
      i. Recall: harvest all
      ii. Technically not feasible
      iii. Large overhead for many purposes
   (b) Restrict scope of crawl to specific subject

2. Approaches
   (a) Subject classification
   (b) Page “importance”, backlink counts
   (c) Depth-First, Breadth-First, Best-First
   (d) Use user information: bookmarks
   (e) Link topology

See Mercator:
http://research.compaq.com/SRC/mercator/research.html
Citation analysis and social networks

1. WWW retrieval problems
   (a) Size:
      i. Low recall rates
      ii. Precision?
   (b) Heterogeneous:
      i. jargon
      ii. junk
      iii. ranking means little
      iv. content matching=bad idea

2. approaches based on social network analysis
   (a) like humans, some pages are more equal than others
   (b) find alpha-male of web pages through social structure
   (c) social network analysis on graph structure of web
Social networks
Social Networks and bibliometrics

Traditional

1. undirected or directed graph
2. edges represent social relationships
   (a) frequency of conversation
   (b) rating of friendship
   (c) phonecalls
   (d) email
   (e) co-authorship

Extended: Bibliometrics

3. structural representation of human relationships
   (a) status
   (b) Positions

1. Any network generated by humans
2. Citation network
3. WWW hyperlinks
Prestige and Power in social networks

<table>
<thead>
<tr>
<th>Prestige and Power</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Each actor is characterized by structure of relationships</td>
<td>1. range of metrics</td>
</tr>
<tr>
<td>2. High v. low number of relationships</td>
<td>2. Degree centrality</td>
</tr>
<tr>
<td>(a) high: high prestige</td>
<td>3. Closeness centrality</td>
</tr>
<tr>
<td>(b) low: low prestige</td>
<td>4. Betweenness centrality</td>
</tr>
<tr>
<td>3. metrics of “centrality”</td>
<td>5. Bonacich centrality</td>
</tr>
</tbody>
</table>
Social networks and centrality

directed graph $G = (V, E)$

set of $n$ actors denoted $V = \{v_0, v_1, \ldots, v_n\}$

$E$ represents the edges between pairs of journals; $E \subseteq V^2$.

Degree centrality

$c_i = \frac{\sum_{j=0}^{n} w_{ij} + \sum_{j=0}^{n} w_{ji}}{\sum_{j=0}^{n} \sum_{k=0}^{n} w_{kj}}$ \hspace{1cm} (1)

Closeness centrality

$d(v_i, v_j)$ is geodesic distance between $v_i$ and $v_j$

$c_i = \frac{\sum_{i=0}^{n-1} d(v_i, v_j)}{n}$ \hspace{1cm} (2)

Betweenness centrality

$g_{ijk}$ number of geodesic paths between any pair of nodes $(v_i, v_j)$ that passes through node $v_k$

$b_k = \frac{g_{ijk}}{g_{ij}}$ \hspace{1cm} (3)
Bonacich centrality

**Recursive definition**

Prestige of actor is function of prestige of actors that related to this actor.

Assume vector $\vec{p}$, entries represent $p(v)$ represent prestige of node $v$.

Adjacency matrix is $E$ so that $E(i, j) = 1$ only when $(v_i, v_j) \in V^2$.

Given a certain initial $\vec{p}$:

$\vec{p}' = E^T \vec{p}$

To determine fixpoint for prestige vector, initialize $\vec{p} = (1, 1, \cdots, 1)$ and iteratively $\vec{p} \leftarrow E^T \vec{p}$ (normalizing values!)

Convergent solution is principal eigenvector.

**Eigenvector centrality**

A little like paths are traversed starting from every possible position. Nodes that are visited most often: highest prestige.
Bibliometrics

Impact Factor

Citation graph over journals.

Impact Factor for journal $i$ in year $x$: $\text{IF}(i,x)$

$F_x(v_i)$ is frequency by which articles in journal $v_i$ are cited in 2 years before $x$.

$n$ number of articles published in journal $i$ over two year period before $x$

$\text{IF}(i,x) = \frac{F_x(v_i)}{n}$

Uses

1. Impact determination of journals
2. Evaluation of publications, research terms, etc.
3. Scientometrics: determining structure of scientific efforts
Applications to WWW

WWW is a directed graph:

1. Determine prestige of pages
2. Rank retrieval results in search engine according to page prestige
3. Only most authoritative pages presented to user

HITS

PageRank

1. Each page on WWW has prestige value
2. Query selects subset of collection
3. Ranked according to query
HITS

1. developed by Jon Kleinberg (Cornell, IBM Almaden)

2. Basic idea:
   (a) Prestige or influence has two flavors
      i. Authorities: high quality information
      ii. Hubs: comprehensive list of links to authorities
   (b) Recursive definitions
      i. Relations to Bonacich or eigenvector centrality
      ii. Every page has two degrees of hubness and authoritativeness
Hubs and Authorities

How is the Web organized? (Left) Web pages can be defined as hubs and authorities. A hub is a page that points to many authorities, whereas an authority is a page that is pointed to by many hubs (71). Characteristic patterns of hubs and authorities can be used to identify communities of pages on the same topic. (Right) An alternate method for identifying communities seeks a set of nodes for which the link density is greater among members than between members and the rest of the network (75).
Authority and Hub scores

**Authority**: sum of hub score of pages linking to it

**Hub**: sum of authority score of pages to which it links

Hub scores over all pages in collection or subset corresponding to query:

\[ \vec{h} \]

Authority scores over all pages in collection or subset corresponding to query:

\[ \vec{a} \]

\[ vech = E\vec{a} \quad (4) \]

\[ veca = E^t\vec{h} \quad (5) \]

Conversion of *vech* and *veca* results in set of authority and hubness values for pages.
HITS procedure

Query first approach

1. Query: text based IR system
2. Expand root set by radius one
3. Power iterations on hubs and authorities
4. Report top ranking hubs and authorities

Topic Distillations

1. No pre-computation of scores
2. As in real-life situations hubs are most useful
3. In WWW: hubs acquire authority: need for hub score?
PageRank

1. Introduced by Brin and Page, Stanford
2. Notion of Random Walk
3. Start user at random positions
   (a) probability of finding users at page X after n clicks
   (b) recursive procedure relating to infinite, random walk
4. principle of eigenvector centrality:
   \[ p_1 = L^t p_0 \]
5. problem:
   (a) Web is not strongly connected
   (b) leaf nodes and cycles

Principle of pre-calculation

1. Once WWW subgraph is know: each page estimated prestige
2. Query isolates documents, ranked according to prestige
PageRank: problems

1. Hyperlink design
   (a) Hyperlink → authority or influence?
   (b) WWW: designed by humans?

2. Nepotism
   (a) Pages on same web site or conglomerate link to each other
       (b) Can warp pagerank

3. Clique attacks
   (a) Deliberate attempt to link clique of pages
   (b) Case of Search King v. Google
Clique attack

**Figure 7.6** How a clique attack takes over link-based rankings.
Webology discussion

![Diagram of the Web as a bow tie](image)

<table>
<thead>
<tr>
<th>Region</th>
<th>SCC</th>
<th>IN</th>
<th>OUT</th>
<th>Tendrils</th>
<th>Disconnected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>56,463,993</td>
<td>43,343,168</td>
<td>43,166,185</td>
<td>43,797,944</td>
<td>16,777,756</td>
<td>203,549,046</td>
</tr>
</tbody>
</table>

**Figure 7.21** The Web as a bow tie [28].