Structure and analysis of www

Rik Sarkar
Hyperlinks

• Give a network structure to a set of documents
  • Instead of being a simple set of documents

• Similar structure in:
  • Citations: articles, patents, legal decision,
    • Usually acyclic: citing only past documents

• Web is more dynamic — pages are updated
  • not acyclic
Connected components

• In a graph:
  • A connected component is a maximal subset of nodes with a path between any pair of nodes in the subset

• In a directed graph (like the web):
  • We are interested in strongly connected components (SCC)
  • An SCC is a maximal subset of nodes, with a directed path between any ordered pair of nodes
    • So, there must be a bath between (a, b)
    • And also between (b, a)
Bow tie structure of the web

Broder '99
Bow tie structure of the web

• *Single* Giant strongly connected component

• Largely due to:
  
  • Many topics are related to each-other (e.g. wikipedia)

  • Many search/directory sites have links to important sites, and these have links to directory/landing sites
Bow tie structure of the web

- *Single* giant SCC
  - hard to have 2 without links between them..

- IN nodes:
  - Flow into the GSCC

- OUT nodes:
  - Flow out of the GSCC

- Structures that do not touch GSCC
  - Tendrils: Flow into OUT and out of IN
  - Tubes: go from IN to out
  - Disconnected pieces
Bow tie structure

• Similar structures in
  • Larger & recent web graphs
  • Wikipedia
• …
Related: Who controls the world?

- The network of global corporate (TNC) control
- Bow tie structure
- The SCC is relatively small
- TNCs in SCC own most of each other
- A group of 147 entities in SCC control About half of World’s economic value
- 3/4 of the SCC are financial intermediaries

S. Vitali et al. 2011
Searching the web

- Search for “Edinburgh” (Information retrieval)
- Find pages that match “Edinburgh”
- Decide which pages are important
Searching the web

• How do you decide:
  • University of Edinburgh is more important than
  • Edinburgh dry-cleaners
• Analyze the web graph to see which node is more important
The basic idea

• In-links constitute a vote for importance
  
  • If somebody is linking to a web page, that means they see something of value in it
  
  • If many people are linking to it, then likely the page is valuable to many other people as well
Enhanced idea

• Not all links imply equal importance

• Links from *Important* pages are more valuable than links from unimportant pages

• Thus, we have an iterative idea:

  1. Decide importance of pages
  2. Update importance of their neighbors suitably
  3. Repeat
The HITS algorithm

• Not all pages are similar

• Some are important for the information they contain (Authorities) (e.g. course pages)

• Some are important for the links they contain (Hubs) (e.g. list of courses)
  • They guide you to the right authorities

• Let’s rank them separately, but depending on each other
  • A hub linking to good authorities is likely good
  • An authority linked by good hubs is likely good
Hubs and authorities

• For each page p, estimate its score both as:
  • A hub: $hub(p)$
  • An authority: $auth(p)$

• Repeatedly in each round
Update rules

• Start with all hub and auth = 1

• Apply Authority update to all nodes:
  • auth(p) = sum of all hub(q) where q -> p is a link

• Apply Hub update to all nodes:
  • hub(p) = sum of all auth(r) where p->r is a link

• Repeat for k rounds
Normalize

- We need only relative values.
- Divide each $auth(p)$ by sum of all $auth$ scores
- Divide each $hub(p)$ by sum of all $hub$ scores
Pagerank

- Idea: Not all pages have good classification as hubs/authorities
- Sometimes authorities link directly to each-other
  - Eg. wikipedia pages
Pagerank: basic algorithm

- Overall “value” in the system is conserved = 1
- Assign “value” 1/n to each node
- In each round
  - Each node divides equal portion of its pagerank value to its out-going links
  - Updates its own value to be sum of values it receives
What are the difficulties of pagerank?
What are the difficulties of pagerank?

- Acyclic graph:
  - Some nodes can get all the values
  - Lakes/seas at the local minima
  - Some nodes can end without any value
  - Rivers or peaks (maxima)
Scaled pagerank

• In every round:
  • Divide $s$ fraction of your pagerank equally among neighbors
  • Divide $(1-s)$ fraction equally among all nodes in the network
The random-walk interpretation

- Users start at random web pages
- Then click links on them randomly
- Sometimes (with Pr = 1-s) they decide to leave the page and jump to a random page in the web
Other improvements

• Use textual information

• Use usage data: which links people click

• Use other contextual data
  • Location, personal history etc…

• Adjustment to SEO

• Adaptation to the fast changing web…
Properties

- HITS converges
- Pagerank Converges
- Pagerank is equivalent to random walk
Before next class

• Please read:

• Chapter 13 & 14 in Kleinberg & Easley

• Including advanced material in ch 14.

• We will cover that in class
Projects

• Will be given end of this week (thursday/friday)
• Deadline nov 25
• Choose one from a set of about 10 to 15
• Each can be taken by at most 5 people
• You can work (discuss) in groups of 1, 2 or 3
• Everyone must submit their own final report and code

• Lookout for email
Adjacency Matrix

Work this out on your own and see if it makes sense:

- $M(i,j) = 1$ iff there is an edge $i \rightarrow j$

- $M(i,j) = 0$ otherwise

Now suppose $a$ is the vector of authority values

Then the hub update rule is equivalent to:

- $h := Ma$