Information Retrieval and Web Search

Link analysis

Instructor: Rada Mihalcea

(Note: This slide set was adapted from an IR course taught by Prof. Chris Manning at Stanford U.)
The Web as a Directed Graph

- **Assumption 1**: A hyperlink is a quality signal.
  - The hyperlink $d_1 \rightarrow d_2$ indicates that $d_1$’s author deems $d_2$ high-quality and relevant.

- **Assumption 2**: The anchor text describes the content of $d_2$.
  - We use anchor text somewhat loosely here for: the text surrounding the hyperlink.
    - Example: “You can find cheap cars `<a href=http://...>here</a>.”
    - Anchor text: “You can find cheap cars here”
[text of $d_2$] only vs. [text of $d_2$] + [anchor text $\rightarrow d_2$]

- Searching on [text of $d_2$] + [anchor text $\rightarrow d_2$] is often more effective than searching on [text of $d_2$] only.

- Example: Query *IBM*
  - Matches IBM’s copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - May not match IBM home page!
    - In particular if the IBM home page contained mostly graphics

- Searching on [anchor text $\rightarrow d_2$] is better for the query *IBM*.
  - In this representation, the page with most occurrences of *IBM* is www.ibm.com
Anchor Text Containing IBM (-> www.ibm.com)

www.nytimes.com: “IBM acquires Webify”

www.slashdot.org: “New IBM optical chip”

www.stanford.edu: “IBM faculty award recipients”

www.ibm.com
Origins of PageRank: Citation Analysis

- Citation analysis: analysis of citations in the scientific literature
- Example citation: “Miller (2001) has shown that physical activity alters the metabolism of estrogens”
  - We can view “Miller (2001)” as a hyperlink linking two scientific articles
- One application of these “hyperlinks” in the scientific literature:
  - Measure the similarity of two articles by the overlap of other articles citing them
  - This is called cocitation similarity
  - Cocitation similarity on the web: Google’s “find pages like this” or “Similar” feature
Another application: Citation frequency can be used to measure the impact of an article
  - Simplest measure: Each article gets one vote – not very accurate
On the web: citation frequency = inlink count
  - A high inlink count does not necessarily mean high quality ...
    - ... mainly because of link spam
Better measure: weighted citation frequency or citation rank
  - An article’s vote is weighted according to its citation impact
  - Circular? No: can be formalized in a well-defined way
Origins of PageRank: Citation Analysis

- Better measure: weighted citation frequency or citation rank
  - This is basically PageRank.
- **PageRank** was invented in the context of citation analysis by Pinsker and Narin in the 1960s
  - Asked: which journals are authoritative?
- We can use the same formal representation for:
  - citations in the scientific literature
  - hyperlinks on the web
- Appropriately weighted citation frequency is an excellent measure of quality:
  - both for web pages and for scientific publications.
Early link analysis: simple popularity ordering
Use link counts as simple measures of popularity
Two basic suggestions:
  Undirected popularity:
    Each page gets a score = the number of in-links plus the number of out-links (3+2=5)
  Directed popularity:
    Score of a page = number of its in-links (3)
How do you spam these two heuristics?
PageRank Scoring

- Imagine a browser doing a random walk on web pages:
  - Start at a random page
  - At each step, go out of the current page along one of the links on that page, equiprobably
- In the “steady state” each page has a long-term visit rate - use this as the page’s score
Is it always possible to follow directed edges in the web graph from any node to any other node? Why or why not?
Not Quite Enough

- The web is full of dead-ends
  - Random walk can get stuck in dead-ends
  - When that happens, it makes no sense to talk about visit rates
Teleportation

- At a dead end, jump to a random web page
- At any non-dead end, with probability 15%, jump to a random web page
  - With remaining probability (85%), go out on a random link
  - 15% - a parameter
- Now cannot get stuck locally.
- There is a long-term rate at which any page is visited - how do we compute this visit rate?
Random Walk Algorithms

- Graph centrality algorithm
- Decide the importance of a vertex within a graph
- A link between two vertices = a vote
  - Vertex A links to vertex B = vertex A “votes” for vertex B
  - Iterative voting → Ranking over all vertices
Random Walk Algorithms

- Model a random walk on the graph
  - A walker takes random steps
  - Converges to a stationary distribution of probabilities
    - Probability of finding the walker at a certain vertex
PageRank
PageRank

A

B

C

D

E

F

0.36

0.32

0.24

0.59

0.22

0.82

0.36

0.22

0.59

0.24
PageRank
PageRank

1.18  0.44  0.33  1.06  0.57  1.34  0.33
PageRank

A → B: 0.51
B → D: 1.35
C → B: 0.63
D → E: 1.17
D → F: 1.47
E → A: 0.36
F → E: 1.47
PageRank

- Usually applied on directed graphs
  - From a given vertex, the walker selects at random one of the out-edges
- Given $G = (V, E)$ a directed graph with vertices $V$ and edges $E$
  - $\text{In}(V_i) =$ predecessors of $V_i$
  - $\text{Out}(V_i) =$ successors of $V_i$

$$S(V_i) = \frac{(1 - d)}{N} + d \sum_{j \in \text{In}(V_i)} \frac{1}{|\text{Out}(V_j)|} S(V_j)$$

d – damping factor $\in [0,1]$ (usually 0.85-0.90)
Example

- Assume a Web of 5 pages
- A links to B, C and E
- B links to C
- C links to E
- D links to A and C
- E links to B and D

What is the PageRank score for each of these pages after 2 iterations?
PageRank Summary

- Preprocessing:
  - Given graph of links, compute the PageRank score
- Query processing:
  - Retrieve pages meeting query.
  - Rank them by their pagerank.
  - Order is query-independent.
- The reality:
  - Pagerank is used in Google, but so are many other clever heuristics.
Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find **two** sets of inter-related pages:
  - **Hub pages** are good lists of links on a subject.
    - e.g., “Bob’s list of cancer-related links”
  - **Authority pages** occur recurrently on good hubs for the subject
- Best suited for “broad topic” queries rather than for page-finding queries
- Gets at a broader slice of common **opinion**
Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic
- A good authority page for a topic is *pointed* to by many good hubs for that topic
- Circular definition - will turn this into an iterative computation
An Example

Hubs → Alice → AT&T → Authorities

Hubs → Bob → Sprint → MCI

Long distance telephone companies
Base Set

- Given text query (say *browser*), use a text index to get all pages containing *browser*.
  - Call this the **root set** of pages
  - Root set typically has 200-1000 nodes
- 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**
  - Base set may have up to 5000 nodes
Visualization

Root set

Base set
Distilling Hubs and Authorities

- Compute, for each page $x$ in the base set, a hub score $h(x)$ and an authority score $a(x)$.
- Initialize: for all $x$, $h_{\text{hits}}(x) \leftarrow 1; a_{\text{hits}}(x) \leftarrow 1$;
- Iteratively update:

$$
HITS_A(V_i) = \sum_{V_j \in \text{In}(V_i)} HITS_H(V_j)
$$

$$
HITS_H(V_i) = \sum_{V_j \in \text{Out}(V_i)} HITS_A(V_j)
$$

- Normalization after each step is also needed to ensure convergence
After several iterations
  - Output pages with highest $hits_h$ scores as top hubs
  - Highest $hits_a$ scores as top authorities