Class 9, March 30th
New Style Web Searching

Quality

- Varies widely
- Relevance: not enough
- Value of page = high quality + relevance
Quality Ranking

- Goal: order the answer to a query in decreasing order of value
  - Query independent: assign an intrinsic value to a document independently of the query. E.g., length, vocabulary, publication date, number of citations (in-degree), who publishes it, etc.
  - Query dependent: value is determined with respect to a particular query. E.g. dot product
  - Mixed

Connectivity Analysis

- Mine hyperlink info on the Web
- Assumption:
  - Links often connect related pages
  - Links between pages is a recommendation
    - E.g., co-citation matrix
- Graph:
  - Pages - nodes
  - Link - directed arcs \((u,v)\) Page \(u\) contains hyperlink to \(v\)
Web Search - “Traditional”

- Web is seen as a graph
- Links (as established by `<A HREF= "url"> text </A>`) are important
- Two independent but similar approaches:
  - PageRank: the basis for Google by Brin and Page (Stanford students)
  - HITS: Kleinberg (MIT, IBM, and Cornell)

Web Search: “New Style”

- Web is seen as a graph
- Links (as established by `<A HREF= "url"> text </A>`) are important
- Two independent but similar approaches:
  - PageRank: the basis for Google by Brin and Page (Stanford students)
  - HITS: Kleinberg (MIT, IBM, and Cornell)
Web Search - “New style”

List of URLs

Page Rank

List of modified URLs

Query

Vector Based Model

word count
weights+type
hit

Page Rank: Back links/examples

A

B

C

100

50

53

3

g

50

3

3
Page Rank (2) - Example

PageRank

- Adds an *objective* measure of citation importance to backlinks that reference a site
- Can be described as model of user behavior
  - Assume “random surfer” given a web page at random and keeps clicking on links
  - User never hits back and eventually gets bored and requests another random page
  - Probability that the surfer will visit a page is PageRank (PR)
  - Probability that the surfer will request a random page is (1-d)
• PageRank is defined as follows:

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1 (typically 0.85). Also, C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

\[
PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right)
\]

PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

• Pages have a high PageRank if they have a lot of sites pointing to them
• Pages can also have a high PageRank if they are pointed to by a small number of sites with high PageRanks (e.g. Yahoo!)
Other Components: Anchors

- Often more accurate description of page than page itself
- Anchors exist for pages that cannot be indexed by text-based search engine
- Potential problem of referencing sites that do not exist

Other Components: Type of Hits

- Consider each hit to be one of type: {title, anchor, URL, plain text large font, plain text small font, ...}, each of which has its own type-weight
- Weight score is combined with PageRank and with a scheme similar to Tf-Idf to give a final rank to the document.