What Information can Hyperlinks Convey?

**Isolated documents:**
- Retrieval methods have focused on the content of the document
- Information is provided by the author itself

**Hypertext and citations:**
- Other documents make statement about document
- Structural organization of collection shared among community

**Possible uses:**
- Relatedness of documents
- Centrality of documents
- Authority / prestige of documents
- Social network behind information network
- ...
WWW Pages Classify Other WWW Pages

Hypertext Structure Gives Info about Relatedness!

Idea:
- two pages are similar (with respect to some aspect), if they are frequently co-cited
- the more frequently two pages are co-cited, the more similar they are

Algorithm for finding similar pages:
- use Google to find all pages that link to a given set of pages
- download those pages and count the frequency of their links
Experiment: Human vs. WebLearn

Fill in the missing page:

Hypertext Structure as a Measure of Similarity

- co-citation groups pages by some aspect of similarity
  - aspect not necessarily easy to identify automatically
  - noise: “best viewed with internet explorer”
  - not all aspects of similarity on the WWW
- Bibliometrics [Small, 1973]
- Use for finding related WWW pages [Joachims et al. 1995/1997]
  [Larson, 1996], [Dean & Henzinger, 1999], Commercial: Netscape,
  Google, etc.
- Use in text classification [Chakrabarti et al., 1998], [Joachims et al.
  2001]
Matching User Expectations in Text Retrieval

**Problem:** Many pages match the word “university”, but what are the most important (most popular) pages on this topic?

**In-Links as an Endorsement**

**Approach 1:**
- A document is more important / popular, the more in-links (backlinks) it has.
Simplified PageRank

**Approach 2:**
A document is more important, if it is linked to from many important documents.
- \( u, v \): documents
- \( F(u) \): outlinks out of \( u \)
- \( B(u) \): inlinks pointing to \( u \)
- \( r(u) \): importance of \( u \)

\[
r_{i+1}(u) = \tilde{a} \sum_{v \in B(u)} \frac{r_i(v)}{|F(v)|}
\]

**Iteration of Simplified Page Rank**

What is the problem with this simplified algorithm?
**PageRank**

A document is more important, if it is linked to from many important documents + some smoothing.

- $u, v$: documents
- $F(u)$: outlinks out of $u$
- $B(u)$: inlinks pointing to $u$
- $e(u)$: inherent importance of $u$ (sum to 1)
- $d$: trade-off parameter
- $r(u)$: importance of $u$

$$r_{t+1}(u) = (1 - d) \bar{a}_i \frac{r_i(v)}{|F(v)|} + de(u)$$

Normalize $r$ so that $||r||_1 = 1$.

---

**Random Surfer**

**Model:**

- Forever:
  - with high probability, follow a random link on the page.
  - or with low probability (e.g. 15%), jump to a random page.

What is the probability, that the surfer is currently at page $u$?

$=>$ Probability distribution over pages for Markov Random Walk

Good approximation for “probability that user wants to see this page” !?
Searching with PageRank

Retrieval function combines:
- vector space similarity
- weighting of html tags
- proximity of matches
- anchor text
- PageRank

=> trade-off with different weights

Hubs and Authorities

Idea [Kleinberg, 1998]: A good hub points to many authorities, and an authority is pointed to by many good hubs.

\[
\text{hub}(u) = \sum_{v \in \text{outlinks}(u)} \tilde{a}_{uv} \text{authority}(v)
\]

\[
\text{authority}(v) = \sum_{u \in \text{outlinks}(u)} \tilde{a}_{vu} \text{hub}(u)
\]

=> Eigenvectors of \(A' A\) (authority) and \(AA'\) (hub)