DAT630
Link Analysis

Search Engines, Section 4.5

18/10/2016

Krisztian Balog | University of Stavanger

So far…

- Representing document content
  - Term-doc matrix, document vector, TFIDF weighting
- Retrieval models
  - Vector space model, Language models, BM25
- Scoring queries
  - Inverted index, term-at-a-time/doc-at-a-time scoring
- Fielded document representations
  - Mixture of Language Models, BM25F
- Retrieval evaluation

Today

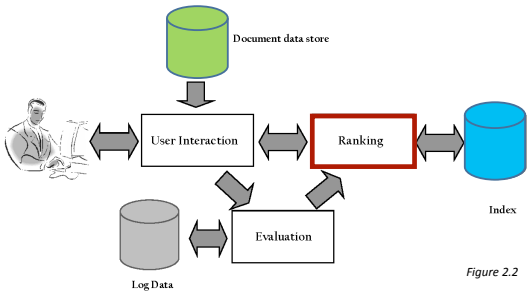
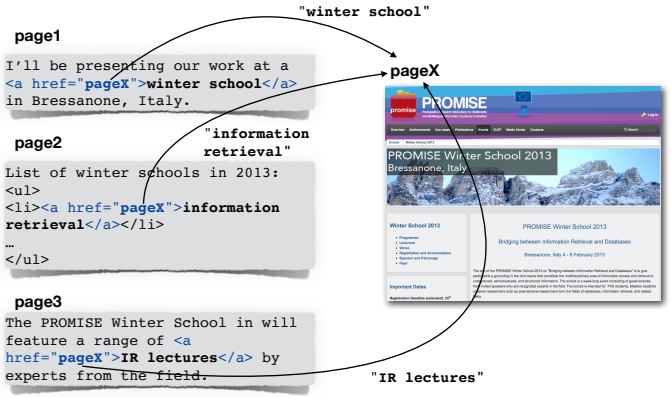


Figure 2.2

Link Analysis

- Links are a key component of the Web
- Important for navigation, but also for search
  - `<a href="http://example.com">Example website</a>`
  - "Example website" is the anchor text
  - "http://example.com" is the destination link
  - both are used by search engines

Anchor Text



Fielded Document Representation

```
title:     Winter School 2013

meta:      PROMISE, school, PhD, IR, DB, [...]
           PROMISE Winter School 2013, [...]

headings:  PROMISE Winter School 2013
           Bridging between Information Retrieval and Databases
           Bressanone, Italy 4 - 8 February 2013

body:      The aim of the PROMISE Winter School 2013 on "Bridging between
           Information Retrieval and Databases" is to give participants a
           grounding in the core topics that constitute the multidisciplinary
           area of information access and retrieval to unstructured,
           semistructured, and structured information. The school is a week-
           long event consisting of guest lectures from invited speakers who
           are recognized experts in the field. [...]

anchors:   winter school
           information retrieval
           IR lectures
```

Anchor text is added as a separate document field

Incorporating Document Importance

$$score'(d,q) = score(d) \cdot score(d,q)$$

Query-independent score
"Static" document score

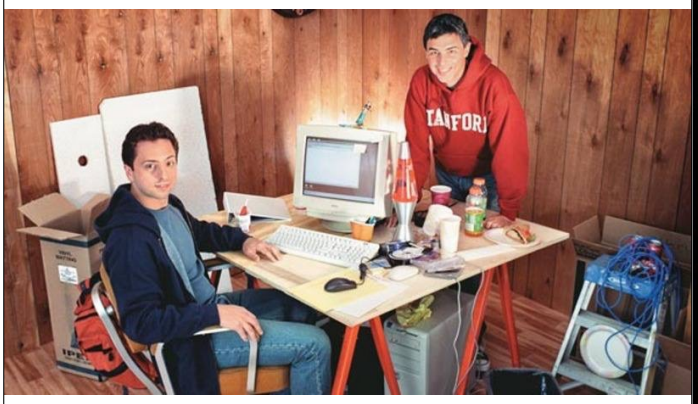
Query-dependent score
"Dynamic" document score

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

Document prior

Document Importance on the Web

- What are web pages that are popular and useful to *many* people?
- Use the links between web pages as a way to measure popularity
- The most obvious measure is to count the number of *inlinks*
  - Quite effective, but very susceptible to SPAM



Google

Search Stanford

10 results | clustering on | Search

Search The Web

10 results | clustering on | Search

Google! BETA

Search the web using Google!

Google Search | I'm feeling lucky

Special Searches
Stanford Search
Linux Search

Help!
About Google!
Company Info
Google! Logos

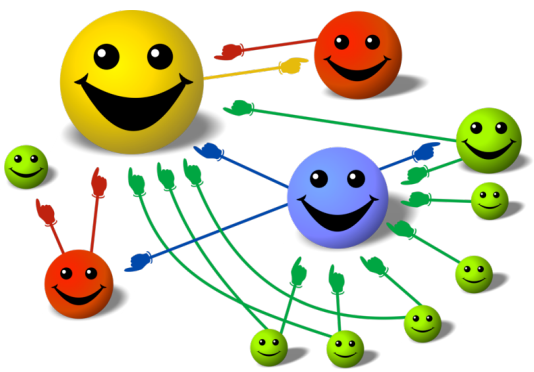
Get Google! updates monthly:
your e-mail
Subscribe | Archive

PageRank

- Algorithm to rank web pages by popularity
- Proposed by Google founders Sergey Brin and Larry Page in 1998
- Thesis: **A web page is important if it is pointed to by other important web pages**

PageRank

- PageRank is a numeric value that represents the importance of a page present on the web
- When one page links to another page, it is effectively casting a vote for the other page
- More votes implies more importance
- Importance of each vote is taken into account when a page's PageRank is calculated



Random Surfer Model

- PageRank simulates a user navigating on the Web randomly as follows:
- The user is currently at page **a**
  - She moves to one of the pages linked from a with probability 1-q
  - She jumps to a random webpage with probability q
- Repeat the process for the page she moved to

This is to ensure that the user doesn't "get stuck" on any given page (e.g., on a page with no outlinks)

PageRank Formula

Jump to a random page with this probability (q is typically set to 0.15)

Follow one of the hyperlinks in the current page with this probability

PageRank value of page p_i

$$PR(a) = \frac{q}{T} + (1-q)\sum_{i=1}^{n}\frac{PR(p_i)}{L(p_i)}$$

PageRank of page a

Total number of pages in the Web graph

Number of outgoing links of page p_i

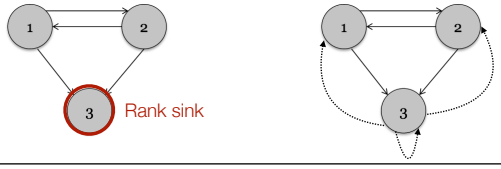
page a is pointed by pages $p_1 \ldots p_n$

Technical Issues

- This is a recursive formula. PageRank values need to be computed iteratively
  - We don't know the PageRank values at start. We can assume equal values (1/T)
- Number of iterations?
  - Good approximation already after a small number of iterations; stop when change in absolute values is below a given threshold

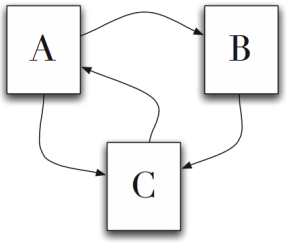
Technical Issues

- Handling "dead ends" (or *rank sinks*), i.e., pages that have no outlinks
  - Assume that it links to all other pages in the collection (including itself) when computing PageRank scores



Example

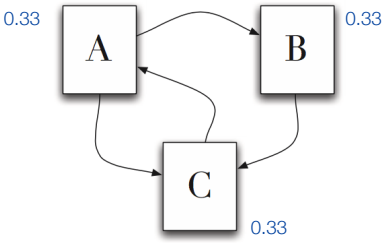
q=0
(no random jumps)



Example

Iteration 0: assume that the PageRank values are the same for all pages

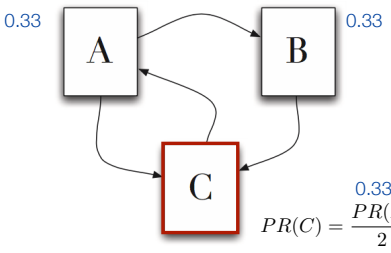
q=0
(no random jumps)



Example

Iteration 1

q=0
(no random jumps)



$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$$

0.33 0.33

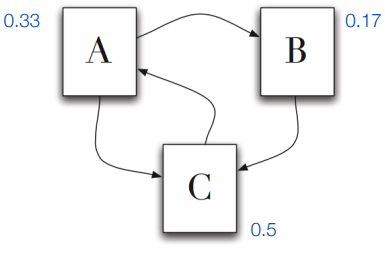
PageRank of C depends on the PageRank values of A and B

=0.5

Example

at the end of **Iteration 1**

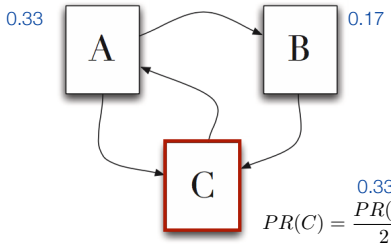
q=0
(no random jumps)



Example

Iteration 2

q=0
(no random jumps)



$$PR(C) = \frac{0.33}{PR(A)} + \frac{0.17}{PR(B)}$$

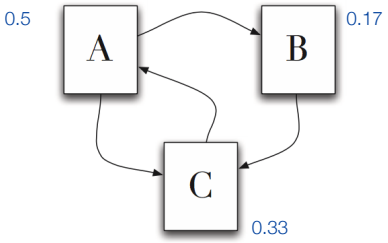
PageRank of C depends on the PageRank values of A and B

=0.33

Example

at the end of **Iteration 2**

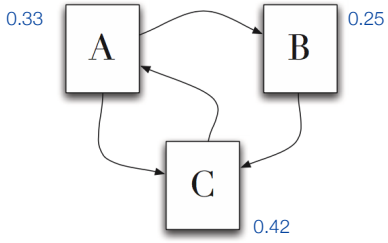
q=0
(no random jumps)



Example

at the end of **Iteration 3**

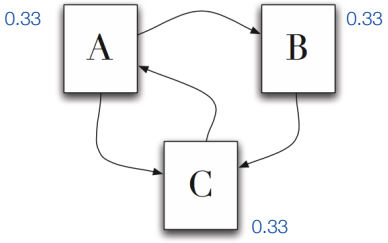
q=0
(no random jumps)



Example #2

Iteration 0: assume that the PageRank values are the same for all pages

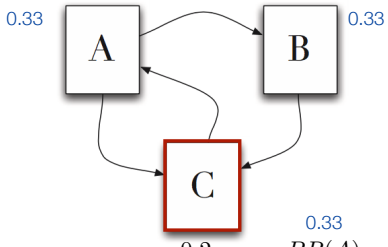
q=0.2
(with random jumps)



Example #2

Iteration 1

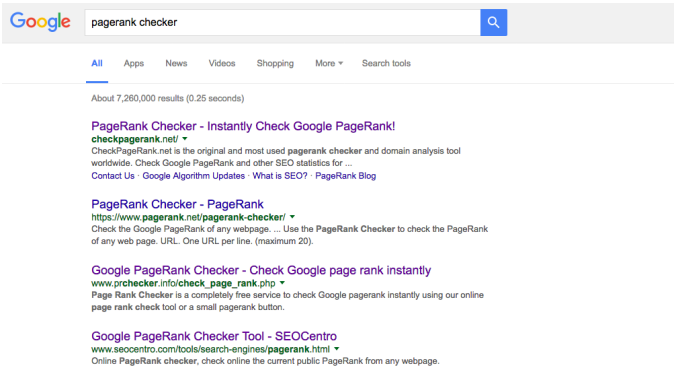
q=0.2
(with random jumps)



$$PR(C) = \frac{0.2}{3} + 0.8(\frac{0.33}{PR(A)} + \frac{0.33}{PR(B)}) \text{ =0.47}$$

Exercise

Online PageRank Checkers



PageRank Summary

- Important example of query-independent document ranking
  - Web pages with high PageRank are preferred
- It is, however, not as important as the conventional wisdom holds
  - Just one of the many features a modern web search engine uses
  - But it tends to have the most impact on popular queries