DAT630
Web Search

Search Engines, Sections 3.2, 4.5

23/10/2017

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

Web search

- Before the web: search was small scale, usually focused on libraries
- Web search is a major application that everyone cares about
- Challenges
  - Scalability (users as well as content)
  - Ensure high-quality results (fighting SPAM)
  - Dynamic nature (constantly changing content)

Some specific techniques

- Crawling
  - Focused crawling
  - Deep web crawling
- Indexing
  - Parallel indexing based on MapReduce
- Retrieval
  - SPAM detection
  - Link analysis

Web Crawling

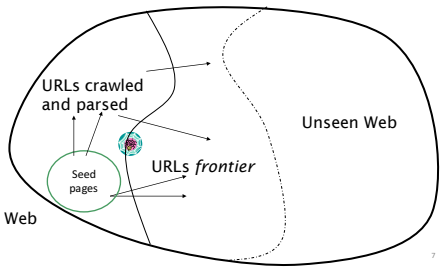
Web Crawling

- Finds and downloads web pages automatically
  - I.e., provides the collection for searching
- Web is huge and constantly growing
- Web is not under the control of search engine providers
- Web pages are constantly changing
- Crawlers also used for other types of data

Web Crawler

- Starts with a set of *seeds*, which are a set of URLs given to it as parameters
- Seeds are added to a URL request queue
- Crawler starts fetching pages from the request queue
- Downloaded pages are parsed to find link tags that might contain other useful URLs to fetch
- New URLs added to the crawler's request queue, or frontier
- Continue until no more new URLs or disk full

Crawling Picture



Web Crawling

- Web crawlers spend a lot of time waiting for responses to requests
- To reduce this inefficiency, web crawlers use threads and fetch hundreds of pages at once
- Crawlers could potentially flood sites with requests for pages
- To avoid this problem, web crawlers use politeness policies
  - e.g., delay between requests to same web server

Web Crawling

- Freshness
  - Not possible to constantly check all pages
  - Must check important pages (i.e., visited by many users) and pages that change frequently
- Focused crawling
  - Attempts to download only those pages that are about a particular topic
- Deep Web
  - Sites that are difficult for a crawler to find are collectively referred to as the *deep* (or *hidden*) Web

Deep Web Crawling

- Much larger than conventional Web
- Three broad categories:
  - Private sites
    - no incoming links, or may require log in with a valid account
  - Form results
    - Sites that can be reached only after entering some data into a form
  - Scripted pages
    - Pages that use JavaScript, Flash, or another client-side language to generate links

Surfacing the Deep Web

- Pre-compute all interesting form submissions for each HTML form
- Each form submission corresponds to a distinct URL
- Add URLs for each form submission into search engine index

Link Analysis

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>
```

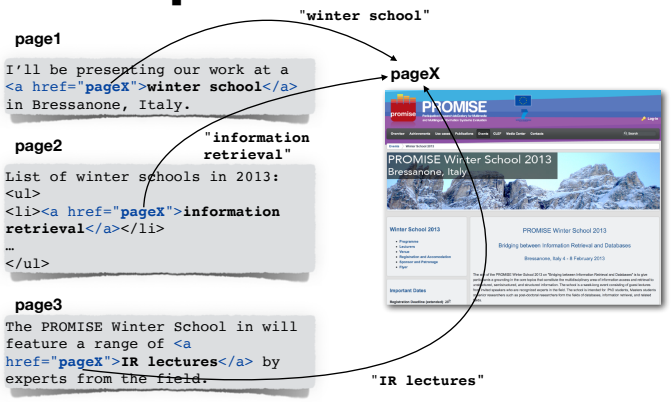
destination link — anchor text

- Both anchor text and links are used by search engines

Anchor text

- Aggregated from all incoming links and added as a separate document field
- Tends to be short, descriptive, and similar to query text
  - Can be thought of a description of the page "written by others"
- Has a significant impact on effectiveness for *some types of queries*

Example



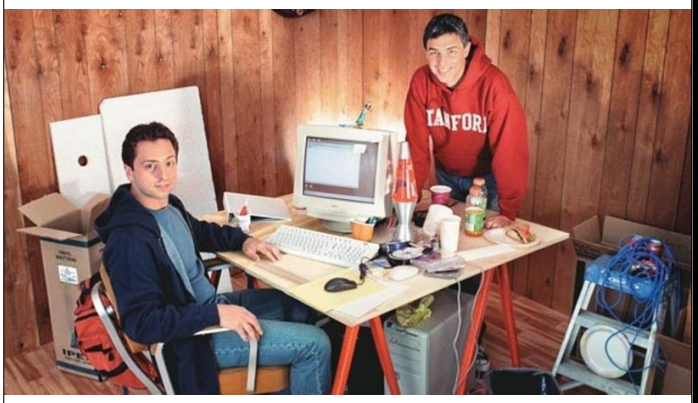
Fielded Document Representation

```
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

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

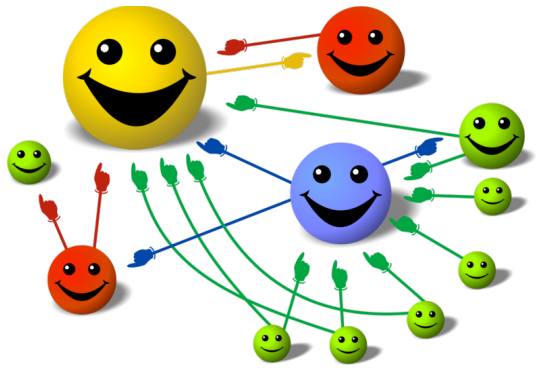


PageRank

- 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

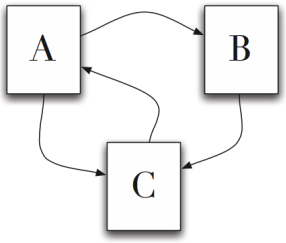
page a is pointed by pages $p_1 \ldots p_n$

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

Example

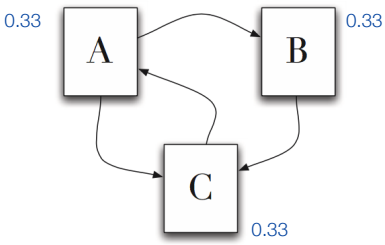
q=0
(no random jumps)



Example

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

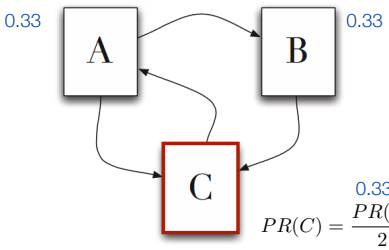
q=0
(no random jumps)



Example

Iteration 1

q=0
(no random jumps)



$$PR(C) = \frac{\overset{0.33}{PR(A)}}{2} + \frac{\overset{0.33}{PR(B)}}{1}$$

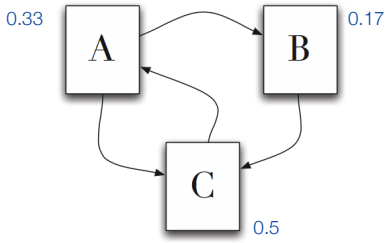
=0.5

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

Example

at the end of **Iteration 1**

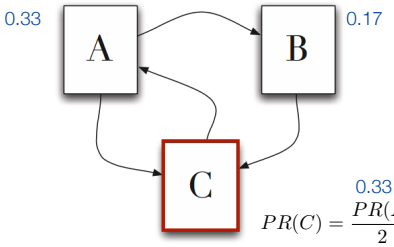
q=0
(no random jumps)



Example

Iteration 2

q=0
(no random jumps)



$$PR(C) = \frac{\overset{0.33}{PR(A)}}{2} + \frac{\overset{0.17}{PR(B)}}{1}$$

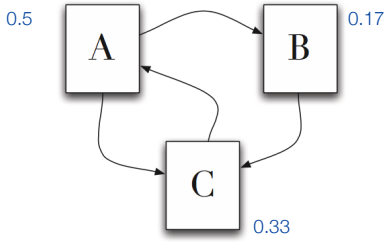
PageRank of C depends on the PageRank values of A and B

=0.33

Example

at the end of **Iteration 2**

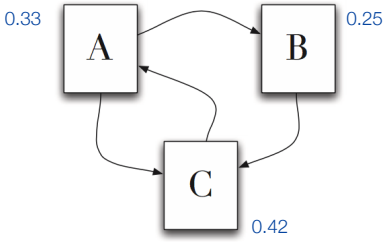
q=0
(no random jumps)



Example

at the end of **Iteration 3**

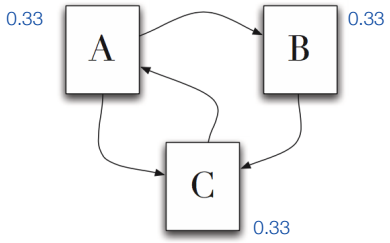
q=0
(no random jumps)



Example #2

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

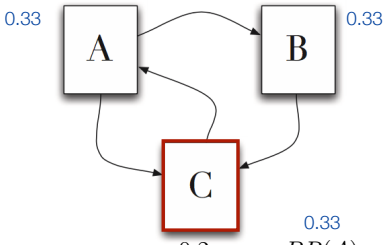
q=0.2
(with random jumps)



Example #2

Iteration 1

q=0.2
(with random jumps)

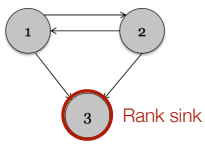


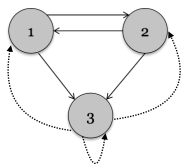
$$PR(C) = \frac{0.2}{3} + 0.8\left(\frac{\overset{0.33}{PR(A)}}{2} + \frac{\overset{0.33}{PR(B)}}{1}\right) \textbf{=0.47}$$

Exercise #1

Dealing with "rank sinks"

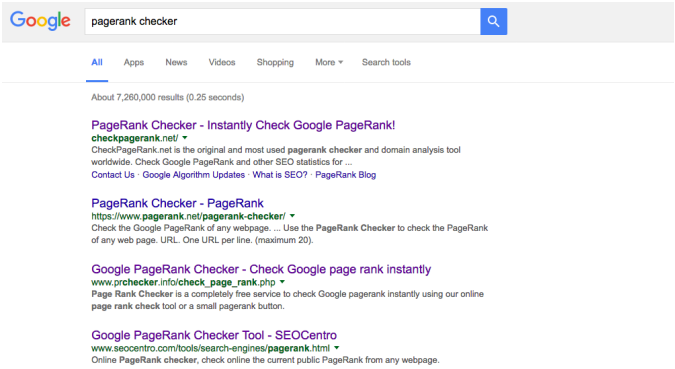
- Handling "dead ends" (or *rank sinks*), i.e., pages that have no outlinks
  - Assume that it links to all other pages in the collection (including itself) when computing PageRank scores





Exercise #2

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

Incorporating Document Importance (e.g. PageRank)

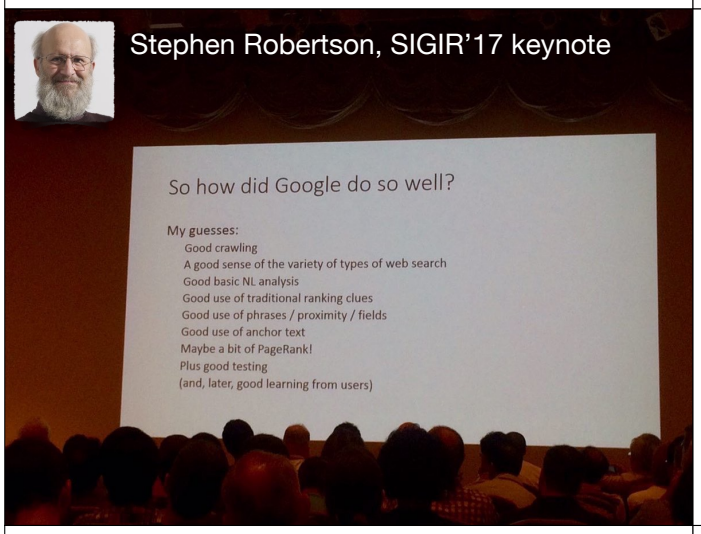
$$score'(d,q) = score(d) \cdot score(d,q)$$

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



Search Engine Optimization

Search Engine Optimization (SEO)

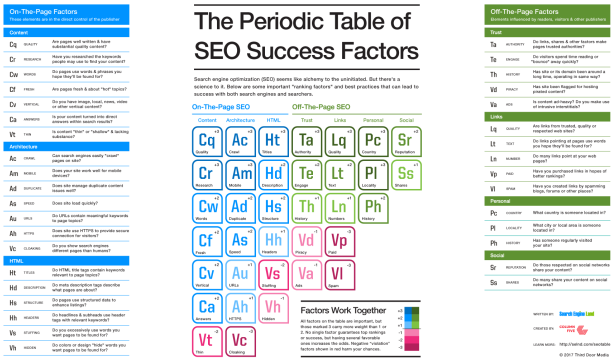
- A process aimed at making the site appear high on the list of (organic) results returned by a search engine
- Considers how search engines work
  - Major search engines provide information and guidelines to help with site optimization
    - Google/Bing Webmaster Tools
  - Common protocols
    - Sitemaps (https://www.sitemaps.org)
    - robots.txt

White hat vs. black hat SEO

- White hat
  - Conforms to the search engines' guidelines and involves no deception
  - "Creating content for users, not for search engines"
- Black hat
  - Disapproved of by search engines, often involve deception
    - Hidden text
    - Cloaking: returning a different page, depending on whether it is requested by a human visitor or a robot

SEO Techniques

- Editing website content and HTML source
- Increase relevance to specific keywords
- Increasing the number of incoming links ("backlinks")
- Focus on long tail queries
- Social media presence



source: http://searchengineland.com/figz/wp-content/seloads/2017/06/2017-SEO_Periodic_Table_1920x1080.png