INFORMATION RETRIEVAL

L10. WEB SEARCH

1

SUZAN VERBERNE 2022

Universiteit Leiden

TODAY'S LECTURE

- The web, web content, web search engines
- Link analysis
- Diversification
- Introduction of final assignment

THE WEB, CONTENT, AND SEARCH ENGINES

IIR CHAPTER 19 & CREDITS TO RICARDO BAEZA-YATES

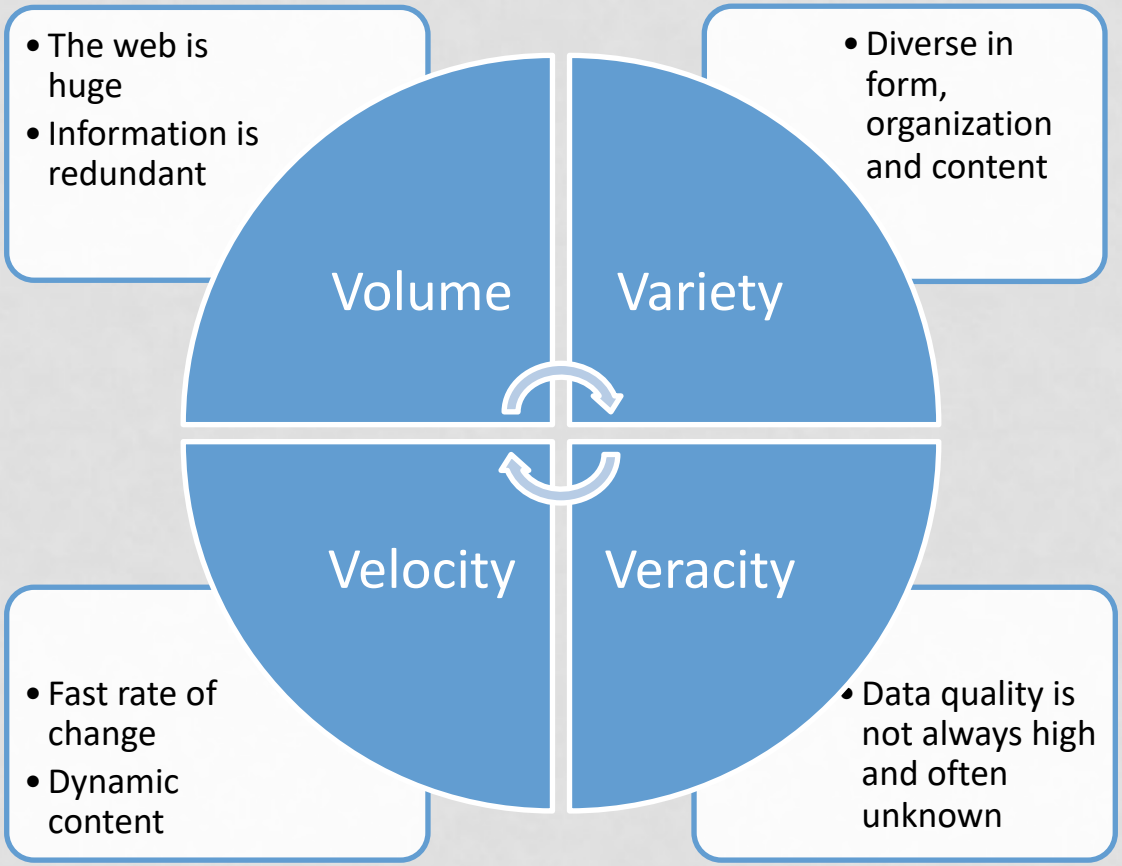
SIZE OF THE WEB

- The web is the largest public repository of data
  - 1.17 billion websites, 17% of these websites are active
  - The indexed web contains at least 3.54 billion pages (2000: 7 million)
  - "As of June 2019, the indexed web was estimated to host 5.85 billion pages — and that's just the activity reached via search engines."

https://siteefy.com/how-many-websites-are-there/
https://www.worldwidewebsize.com/
https://starry.com/blog/inside-the-internet/how-big-is-the-internet

WEB CONTENT

- Web data = big data



WEB CONTENT

- How do we fill the index of a search engine? Which documents are indexed? How do we find them?
- Crawling:
  - Quantity
  - Freshness
  - Quality

conflicting

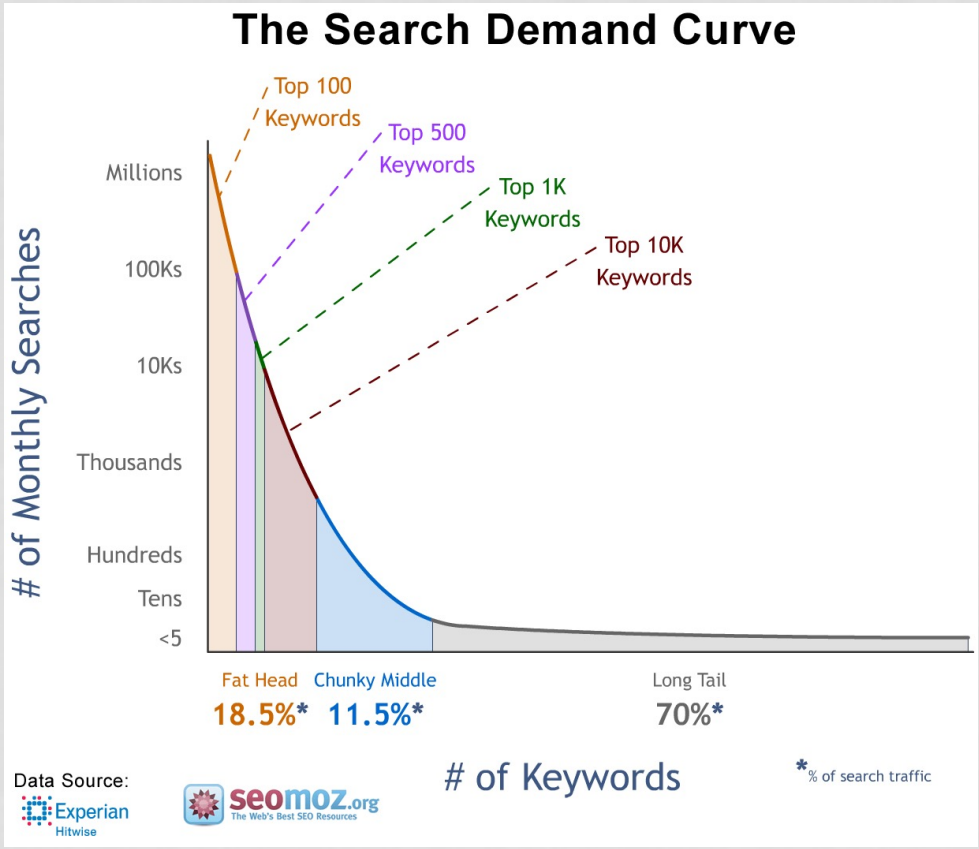
- Strategies for crawling?

CRAWLING

- Breadth-first
- Depth-first
- Importance ordering
  - Sites with highest PageRank scores first
- Size ordering
  - Largest sites first

THE LONG TAIL OF WEB SEARCH

- Heavy tail of user interests
  - Queries issued, documents clicked, domains visited, movies watched, articles read, words used …
  - There are many queries that are asked very few times
  - These make up a large fraction of all queries
- The long tail matters! (because everyone is in there)



https://icenineonline.com/blog/a-simple-explanation-of-long-tail-keywords/

WEB SEARCH ENGINES

The evolution of commercial web search engines

1. First generation: use only text content (1994-1997, e.g. Altavista)
   - Query term frequency (tf-idf, since 1972)
   - Based on library systems from the decades before the web
2. Second generation: use off-page, web-specific data (from 1998, first developed by Google)
   - Link analysis
   - Click-through data (What results people click on)
   - Anchor-text (How people refer to this page)

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/
To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords

World Wide Web, Search Engines, Information Retrieval, PageRank, Google

Universiteit Leiden

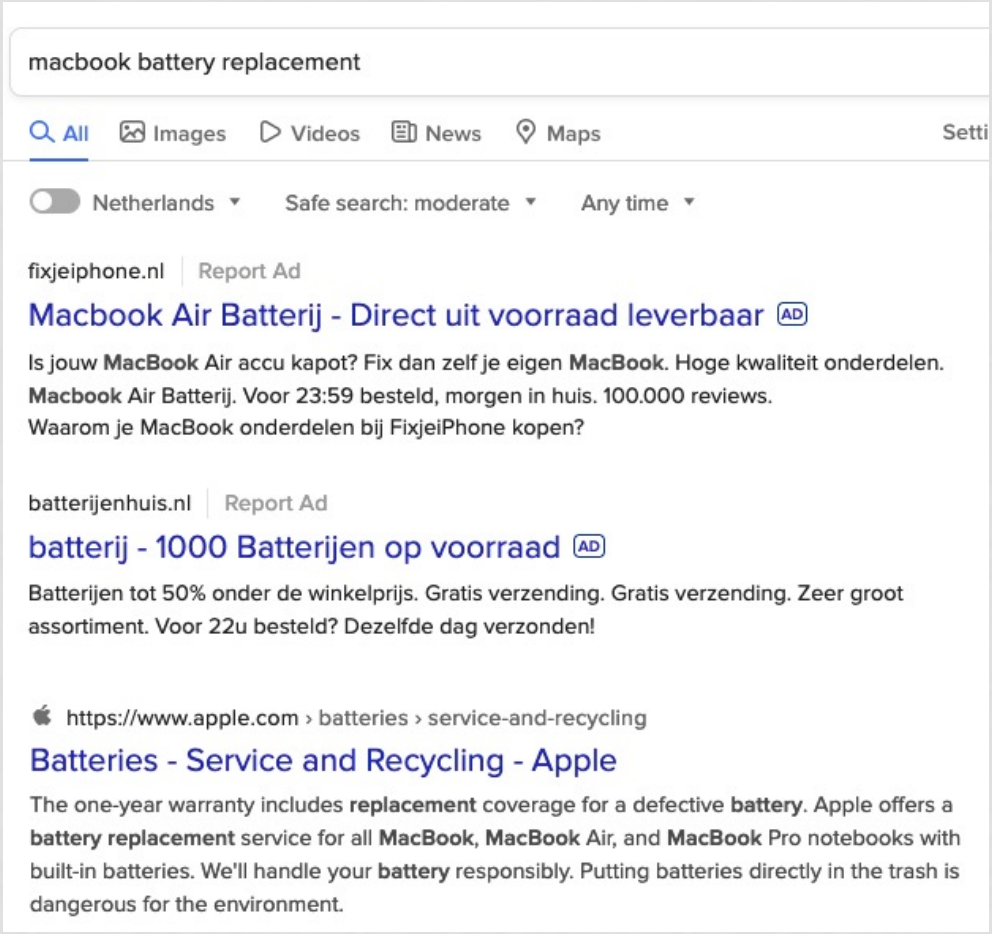
Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117

WEB SEARCH ENGINES

3. Third generation: help the user with their information need
   - Focus on user need, rather than on query text strictly
   - Semantic analysis: what is this about?
   - Integrate multiple sources of data
   - Use of context
     - spatial (user location)
     - query stream (previous queries)
     - personal (user profile)
   - Help the user
     - UI, info boxes, spell checking, query refinement, query suggestion

ADVERTISING

- Users have needs …
- Advertisements are the core business model of web search engines
- Cost per click (CPC) model:
  - Advertisers pay the search engine, get clicks in return
  - Clicking on the advertisement leads to shop/service
  - Goal: induce a transaction



LINK ANALYSIS

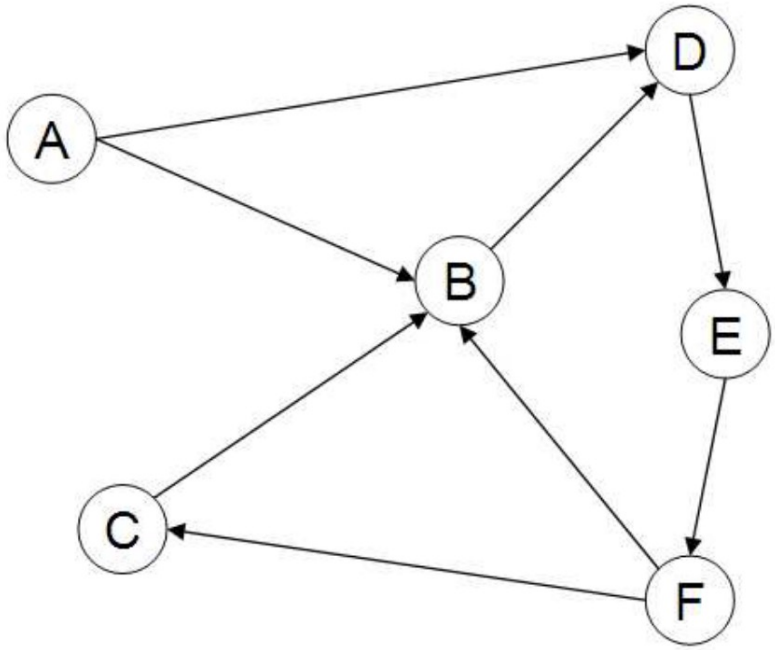
IIR CHAPTER 21

THE WEB AS GRAPH

- Scholarly citations represent evidence of authority from one article to another
- Link analysis on the Web also treats hyperlinks as evidence of authority (or at least an endorsement) from one web page to another as
  - Not every hyperlink implies authority or endorsement
  - Simply measuring the quality of a web page by the number of in-links (hyperlinks from other pages) is not robust enough

THE WEB AS GRAPH

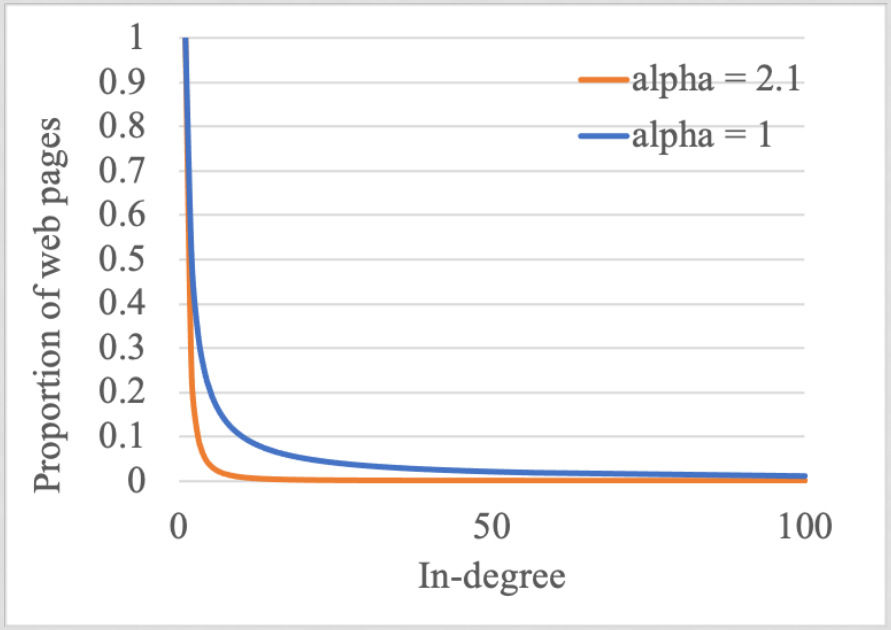
- The (static) Web as directed graph:
  - each web page is a node
  - each hyperlink a directed edge



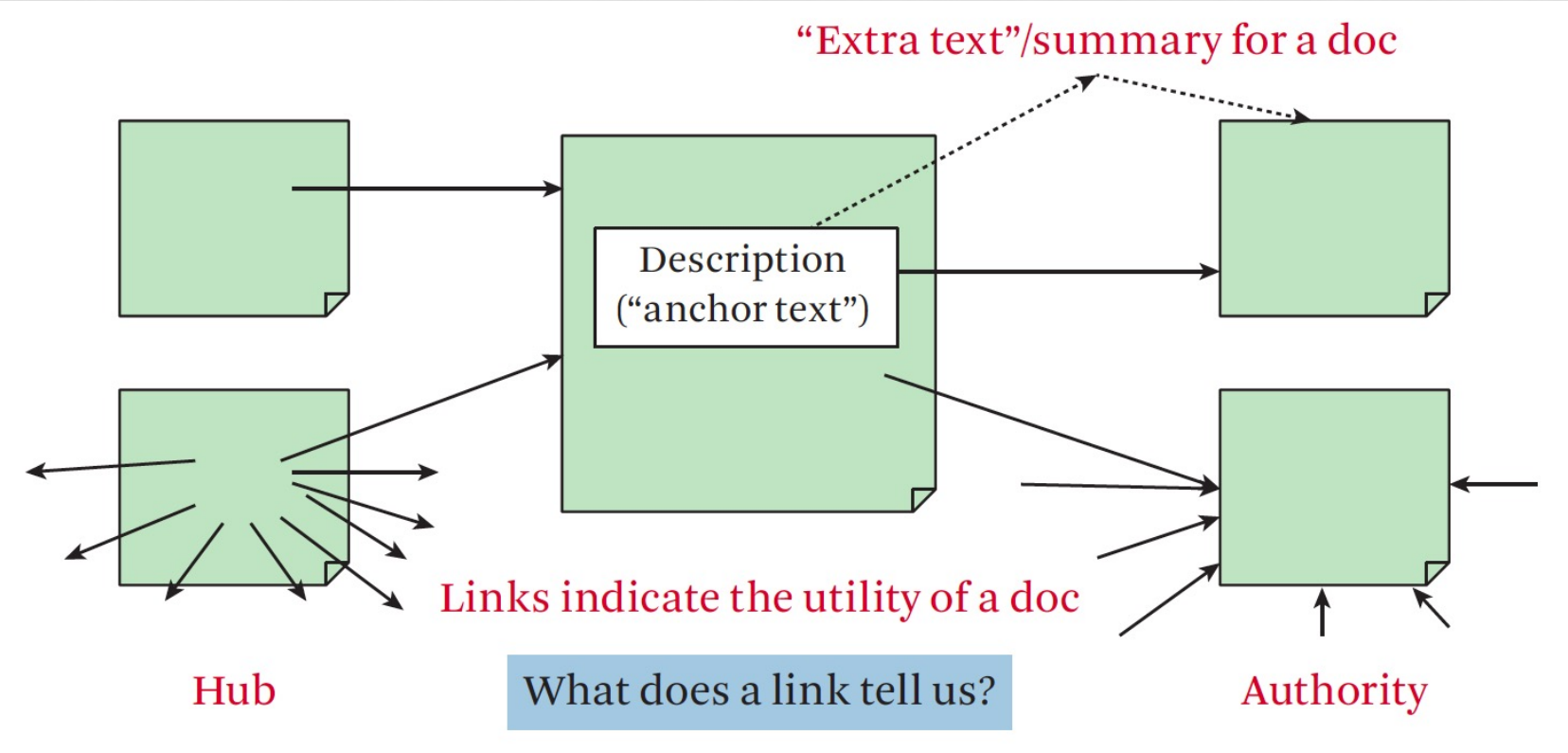
► **Figure 19.3** A sample small web graph. In this example we have six pages labeled A-F. Page B has in-degree 3 and out-degree 1. This example graph is not strongly connected: there is no path from any of pages B-F to page A.

THE WEB AS GRAPH

- The directed graph is not strongly connected
  - there are pairs of pages (a, b) such that b cannot be reached by starting from a and following hyperlinks
- Distribution of number of in-links for a web page is power law ('long tail')
  - number of web pages with in-degree i is proportional to $\frac{1}{i^{\alpha}}$
  - with α estimated as 2.1
  - (Zipf's law for the distribution of words in text is a power law with α = 1)



LINK ANALYSIS



Zhai & Massung 2016

LINK ANALYSIS

- Two intuitions about hyperlinks:
  1. The anchor text pointing to page B is a good description of page B (textual information)
  2. The hyperlink from A to B represents an endorsement of page B, by the creator of page A (quality signal)

- Both signals contain noise

LINK ANALYSIS

- Example of anchor texts:

I am an Associate Professor at the Leiden Institute of Advanced Computer Science (LIACS). I am group leader of Text Mining and Retrieval Leiden.

```
<p>I am an Associate Professor at the Leiden Institute of Advanced
Computer Science (<a href="https://liacs.leidenuniv.nl/" target="_blank"
rel="noopener noreferrer">LIACS</a>). I am group leader of <a
href="http://tmr.liacs.nl/" target="_blank" rel="noopener
noreferrer">Text Mining and Retrieval Leiden</a>.</p>
```

ANCHOR TEXT

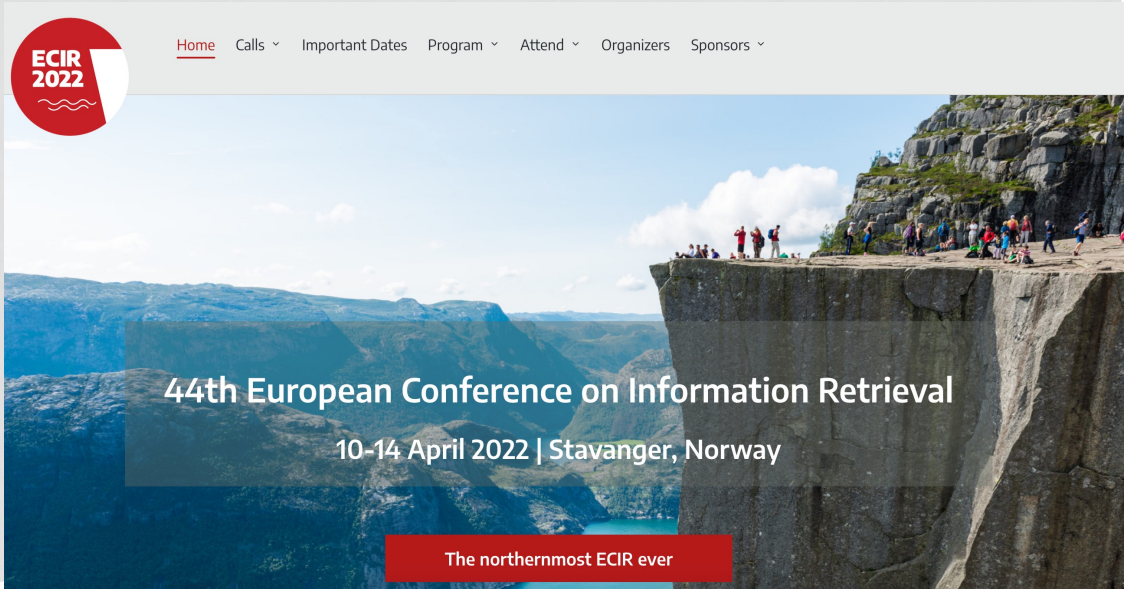
Textual information provided by anchor texts

- Often, the anchor text is descriptive of the URL the hyperlink points to
- This information can be used to bridge the vocabulary gap between query and document: anchor texts may contain query terms that are not in the document itself
  - Example: descriptions/roles of people/organizations
- We can use anchor texts to enrich the indexed information for web pages
  - What to do with meaningless anchor texts (“here”, “this website”)?
  - Weight the terms for their frequency, e.g. using idf

ANCHOR TEXT

Example anchor texts for the same web page:

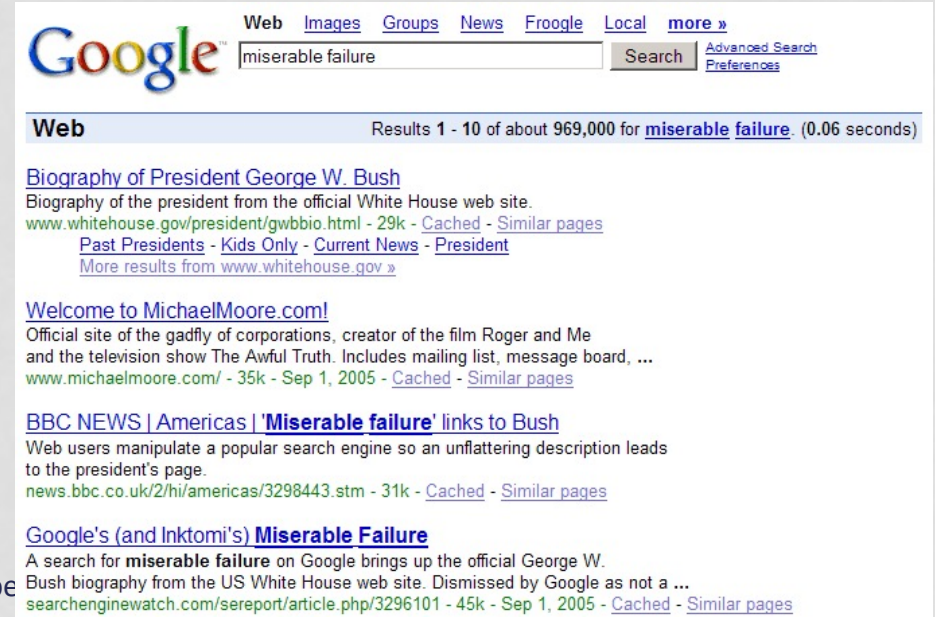
- the 44th European Conference on Information Retrieval
- the European IR conference 2022
- Conference in Stavanger in April
- ECIR Stavanger
- ECIR 2022
- ECIR2022



Universiteit Leiden

ANCHOR TEXT

- Adversarial effect: ‘Google bombs’
  - 1999: “more evil than Satan himself” retrieved the Microsoft homepage as the top result
  - 2006: “miserable failure” retrieved President Bush's biography



LINK ANALYSIS

- Two intuitions about hyperlinks with anchor text:
  1. The anchor text pointing to page B is a good description of page B (textual information)
  2. The hyperlink from A to B represents an endorsement of page B, by the creator of page A (quality signal)

PAGERANK

PAGERANK

- Quality signal provided by hyperlinks: we can derive relevance metrics from the link structure
- Most famous: PageRank

The anatomy of a large-scale hypertextual web search engine

S Brin, L Page - Computer networks and ISDN systems, **1998** - Elsevier

… Another intuitive justification is that a page can have a high **PageRank** if there are many pages that point to it. or if there are some pages that point to it and have a high **PageRank**. …

☆ Save Cite Cited by 21962 Related articles All 222 versions

PAGERANK

- PageRank: technique for link analysis that assigns to every node in the web graph a numerical score between 0 and 1
- The PageRank of a node will depend on the link structure of the web graph
- Main intuition: Pages visited more fequently in a random walk on the web are the more important ones

PAGERANK

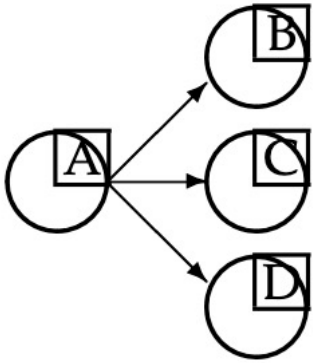
PageRank intuitions:

- Incoming link counts are an important signal
  - A page is useful if it is cited often
- Indirect citations also count
  - If important pages are pointing to a page, that page must be important
- Smoothing
  - Smooth the citations with some random step to accommodate potential citations that have not yet been observed

PAGERANK

- Consider a random surfer who
  - begins at a random web page (a node of the web graph)
  - executes a random walk on the Web
  - At each time step, the surfer proceeds from his current page A to a randomly chosen web page that A hyperlinks to

The figure shows the surfer at a node A, out of which there are three hyperlinks to nodes B, C and D; the surfer proceeds at the next time step to one of these three nodes, with equal probabilities 1/3



PAGERANK

How to estimate PageRank scores:

1. Start at a random page
2. Jump to another page:
   - with probability α to a random page ('teleportation rate')
   - with probability $1 - \alpha$ to any link from the current page
3. Repeat step 2 until convergence
4. PageRank final score for page d: the probability the surfer reaches page d

PAGERANK

$$PR(p) = \frac{\alpha}{N} + (1 - \alpha) \sum_{q \to p} \frac{PR(q)}{O(q)}$$

- N: total number of pages in Web graph
- $PR(q)$: PageRank score of q in current iteration
- $O(q)$: number of outgoing links of page q

In Google PageRank, the factor N is left out

PAGERANK

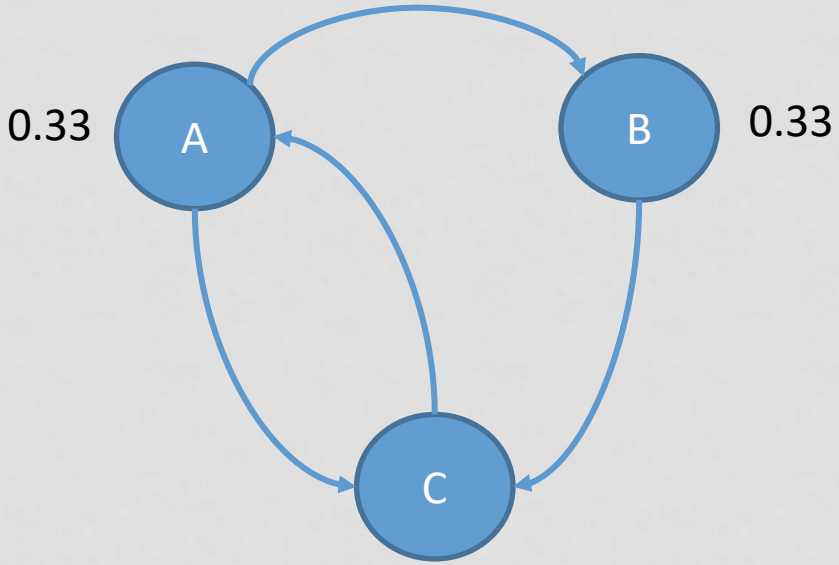
How is PageRank computed?

- Recursive definition but computed iteratively
  - Stops when change in absolute values between 2 iterations is below a threshold
  - The iterative process is guaranteed to converge
- Brin and Page: for a network consisting of 322 million links, PageRank converges in 52 iterations
  - the scaling factor for extremely large networks would be roughly linear in $\log(n)$ where n is the size of the network

PAGERANK

$$PR(p) = \frac{\alpha}{N} + (1-\alpha)\sum_{q \to p} \frac{PR(q)}{O(q)} \qquad \text{with } \alpha = 0$$

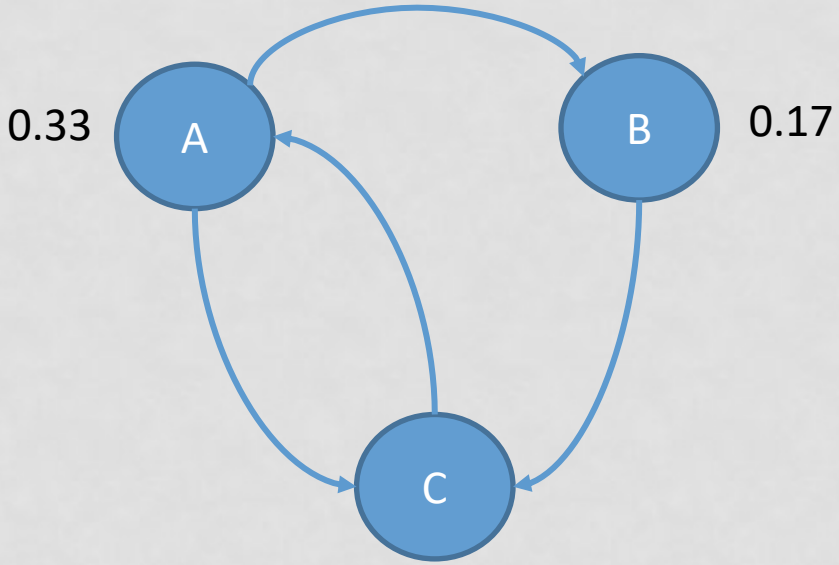
Iteration 0



PAGERANK

$$PR(p) = \frac{\alpha}{N} + (1-\alpha)\sum_{q \to p} \frac{PR(q)}{O(q)} \qquad \text{with } \alpha = 0$$

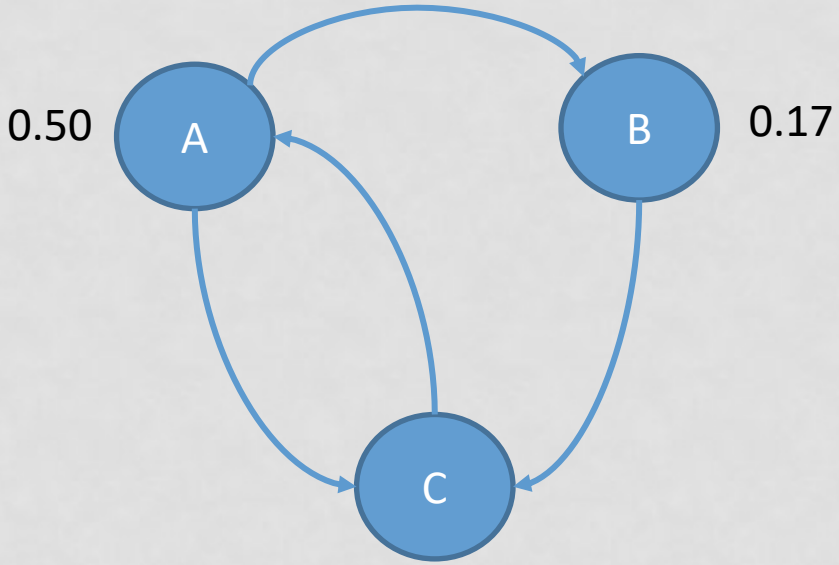
Iteration 1



PAGERANK

$$PR(p) = \frac{\alpha}{N} + (1-\alpha)\sum_{q \to p} \frac{PR(q)}{O(q)} \qquad \text{with } \alpha = 0$$

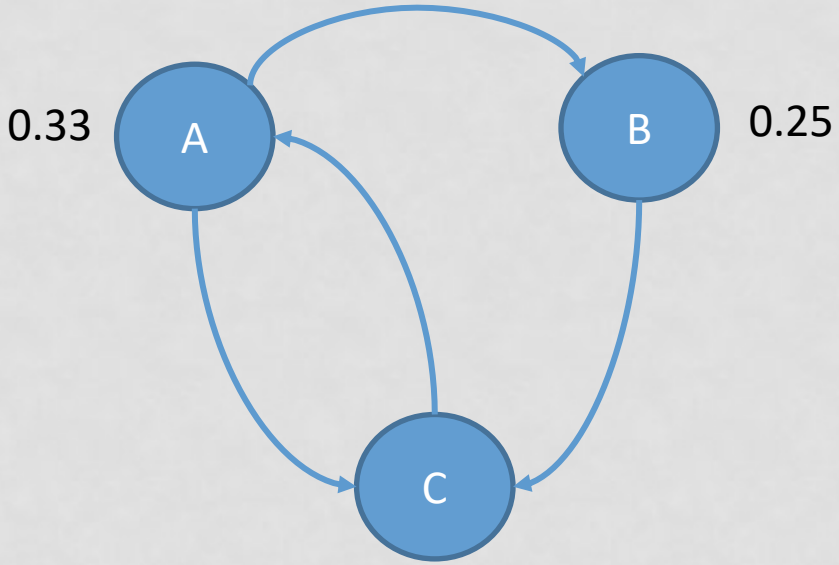
Iteration 2



PAGERANK

$$PR(p) = \frac{\alpha}{N} + (1-\alpha)\sum_{q \to p} \frac{PR(q)}{O(q)} \qquad \text{with } \alpha = 0$$

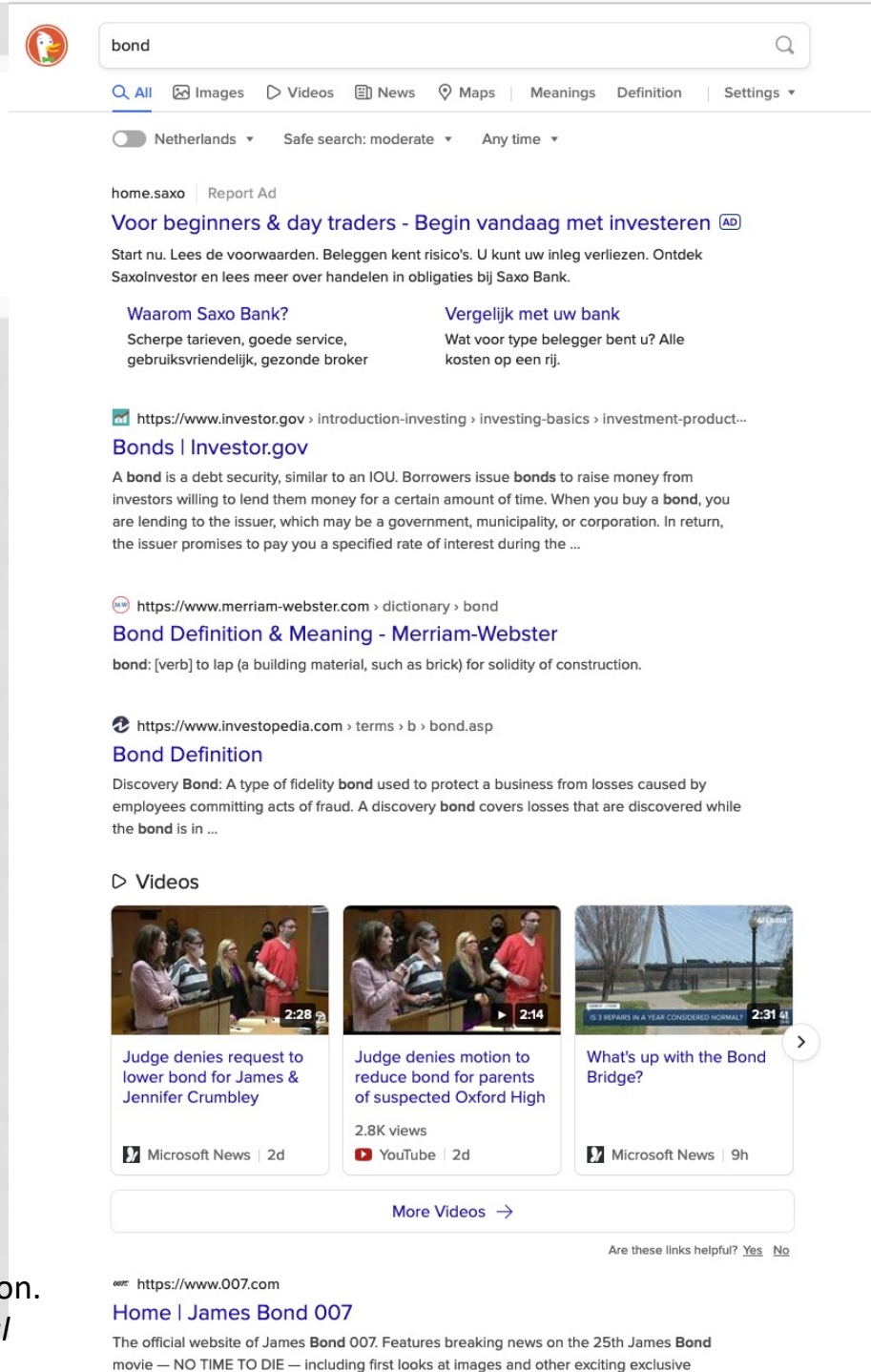
Iteration 3



SEARCH RESULT DIVERSIFICATION

DIVERSITY

- Why diversification?
  1. Queries are often short and ambiguous: we don't know what the user wants
  2. If we take query-document similarity as the most important ranking criterion, there might be a lot of redundance in the top-ranked results



Universiteit Leiden

Rodrygo et al. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*

DIVERSITY

- Diversity-oriented ranking has been proposed as a means to over-come ambiguity and redundancy during the search process
- Principle: not consider the relevance of each document in isolation, but consider how relevant the document is in light of
  A. the multiple possible information needs underlying the query
  B. the other retrieved documents
- Goals:
  - Maximum coverage (of information needs)
  - Minimum redundancy: high novelty of returned documents

IMPLICIT DIVERSIFICATION

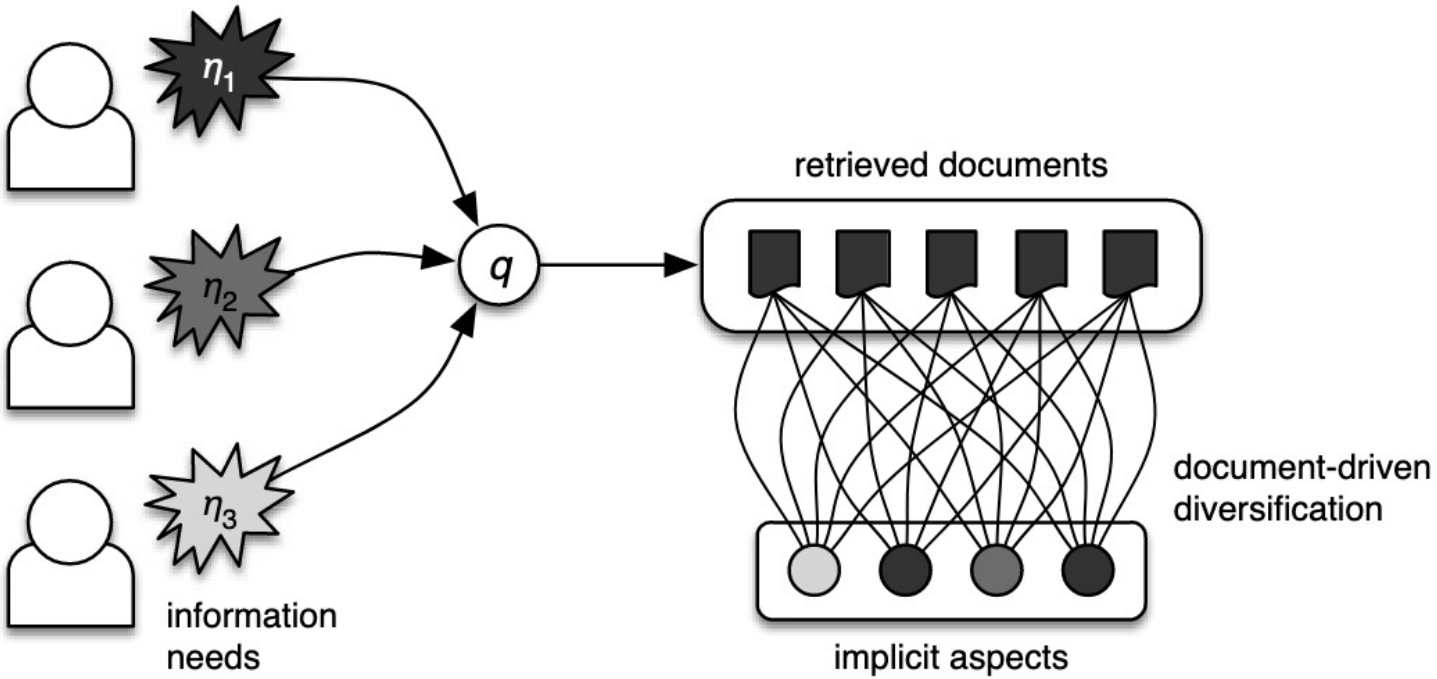


Figure 3.1: Schematic view of implicit diversification approaches.

Rodrygo et al. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*

IMPLICIT DIVERSIFICATION

- Novelty-based diversification approach: maximal marginal relevance (MMR)
  - Score a candidate document $d \in R_q \backslash D_q$ as the document's estimated relevance with respect to the query q, discounted by the document's maximum similarity with respect to the already selected documents in D_q:

$$f_{MRR}(q, d, D_q) = \lambda f_1(q, d) - (1 - \lambda) \max_{d_j \in D_q} f_2(d, d_j)$$

  - $f_1(q, d)$: relevance of d to q
  - $f_2(d, d_j)$: similarity of d_j to d

FINAL ASSIGNMENT

CREDITS TO ARIAN ASKARI

GENERAL INFORMATION

Goals of this assignment

1. to acquire a more in-depth understanding of the effect of modification on scoring functions on retrieval effectiveness
2. to learn to work with ElasticSearch
3. to learn to implement a scoring function based on a mathematical definition

GENERAL INFORMATION

- Group assignment, groups of 3
  - Please enroll in a group to see the assignment in Brightspace
  - Detailed information in the PDF attached to the assignment
  - 20% weight in the course grade
  - Deadline: May 24, 2022

BACKGROUND INFORMATION

- Task: TREC Clinical Trial Track
  - Query by document (QBD) retrieval is a task in which the user enters a text document – instead of few keywords – as a query, and the IR engine finds relevant documents from a text corpus
  - The TREC 2021 Clinical Trials Track is a QBD task focused at finding clinical trials that are eligible for a specific patient
  - There are 375,580 documents, 75 queries and about 11k relevance judgements in the Clinical Trial Track collection

BACKGROUND INFORMATION

- Scoring functions: BM25 variants
  - Perhaps the most well-known IR model
  - Almost always used as first-stage term-matching method before neural re-ranking
  - Many variants of the function exist and the differences matter

BACKGROUND INFORMATION

European Conference on Information Retrieval

↳ ECIR 2020: **Advances in Information Retrieval** pp 28–34 | Cite as

Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants

Chris Kamphuis, Arjen P. de Vries, Leonid Boytsov & Jimmy Lin

Conference paper | First Online: 08 April 2020

5934 Accesses | **9** Citations | **3** Altmetric

Part of the Lecture Notes in Computer Science book series (LNISA,volume 12036)

TASKS

1. Read the paper by Kamphuis et al. (2020): “Which BM25 do you mean? A large-scale reproducibility study of scoring variants.”
2. Read the official website of the TREC Clinical Trials Track to get a holistic view of the IR task of this assignment
3. Install version 7.9.1 of ElasticSearch
4. Download the dataset and index it with ElasticSearch through the Python library
5. Do a first-stage retrieval run for the Clinical Trials Track queries with the default ElasticSearch retrieval function
6. Implement and evaluate two BM25 variants ranking functions for the Clinical Trials Track task

REPORT

7. Write a report in which you
   - motivate the choice for the BM25 variants
   - describe your implementation
   - compare the results
   - discuss the results

- The report's final length should be 3–4 pages (single-column ACM proceeding template or a similar template)
- Deadline: May 24, 2022

HOMEWORK (1)

- Read:
  - IIR book: chapter 21

- Final assignment:
  - Form groups of 3
  - Read the assignment and plan your activities (paper reading, data download, Python and ElasticSearch preliminaries, experimentation, report writing)

HOMEWORK (2)

- Exercises about link analysis
  - Two about anchor texts, one about PageRank
- See Brightspace -> assignments
- Deadline: Sunday May 1, 23.59

AFTER THIS LECTURE…

- You can explain the long tail distribution for multiple aspects of the web and web search
- You can explain the two intuitions about hyperlinks on the web
- You can explain the PageRank algorithm
- You can compute PageRank iteratively for a toy node set
- You can give two reasons for search results diversification
- You can define maximal marginal relevance
- You know what to expect from the final assignment