T-61.6040 Special Course in Computer and Information Science IV: Information Networks

László Kozma (lkozma@cc.hut.fi)

- S.Brin, L.Page: **The Anatomy of a Large-Scale Hypertextual Web Search Engine** (1998)

- L.Page, S.Brin, R.Motwani, T.Winograd: **The PageRank Citation Ranking: Bringing Order to the Web** (1998)

This presentation is about:

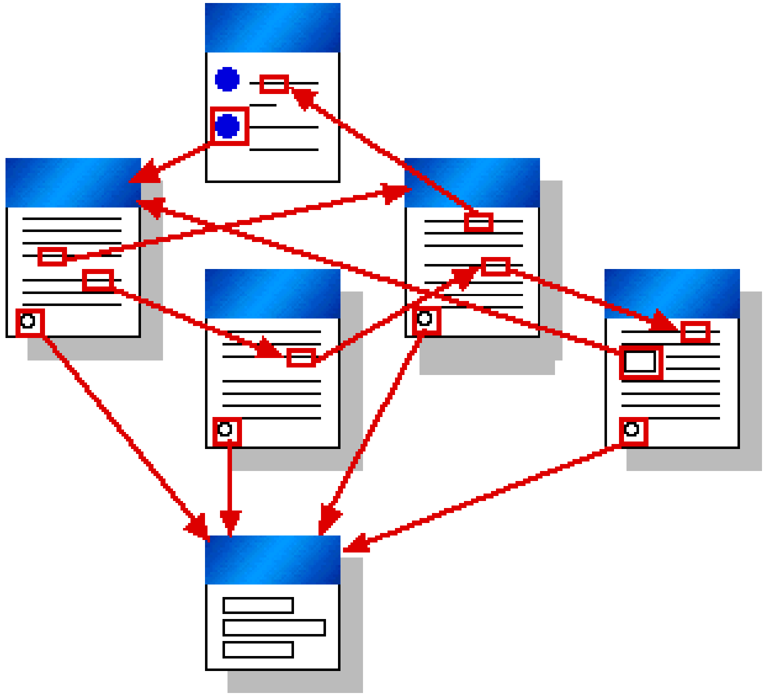
Google web search engine

- Not about:
  - Google Corporation
  - Gmail, Google Maps, Google Docs, etc.
  - NASDAQ:GOOG
  - “Do no evil”, etc. (maybe just a few words)

Web search

- WWW:

  ~30 B pages (source:Netcraft)
- Search engines:
  - crawl
  - index
  - query (by keywords)
  - rank



Web search in 1998

- Not very useful:
  - mostly junk results
  - ranking doesn't work well
  - ex.: 3 of the 4 leading search engines can't find themselves
  - ex.: “Bill Clinton” joke of the day

Web search in 1998

- Design elements:
  - term counting
  - backlink counting
  - meta tags
  - “mixed motives”
  - closed algorithms

How was Google different?

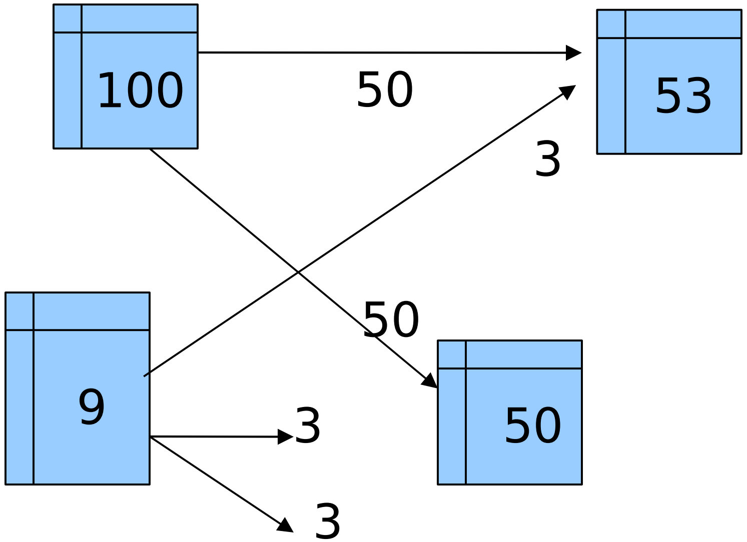
- PageRank
- Using external information
- Scalable Architecture
- Openness, “Scientific Integrity”
- AdSense

PageRank

- “Importance” of a page:
  - is it possible to measure objectively?
- “Academic citations” model
- But the web is different:
  - heterogeneous
  - no quality control
  - ease of publishing
  - manipulation

PageRank

- Basic idea:
  - links are not equally important
  - Assign a Ranking for each page
  - Ranking propagates through links (votes)
  - “votes” evenly distributed among outgoing links



PageRank

- Simplified Ranking

$$R(u) = c \cdot \sum_{v \in B_u} \frac{R(v)}{|F_v|}$$

u – web page

Fu – forward link pages

Bu – backward link pages

PageRank

- “Random Surfer” Model:
  - PageRank as probability distribution
- Problem with previous formula:
  - source sinks
  - Solution: damping factor (random surfer gets bored)
  - Pagerank with damping (typically d=0.85):

$$PR(u) = \frac{1-d}{N} + d \cdot \sum_{v \in B_u} \frac{PR(v)}{|F_v|}$$

PageRank

- Computing PageRank: Iterative approach
- Convergence:
  - Affected by graph structure (good “expansion factor”)
  - Initial values don't affect result, just convergence speed
  - Typically ~log(N) nr. of iterations.
- Variants: personalized PageRank
- Manipulation

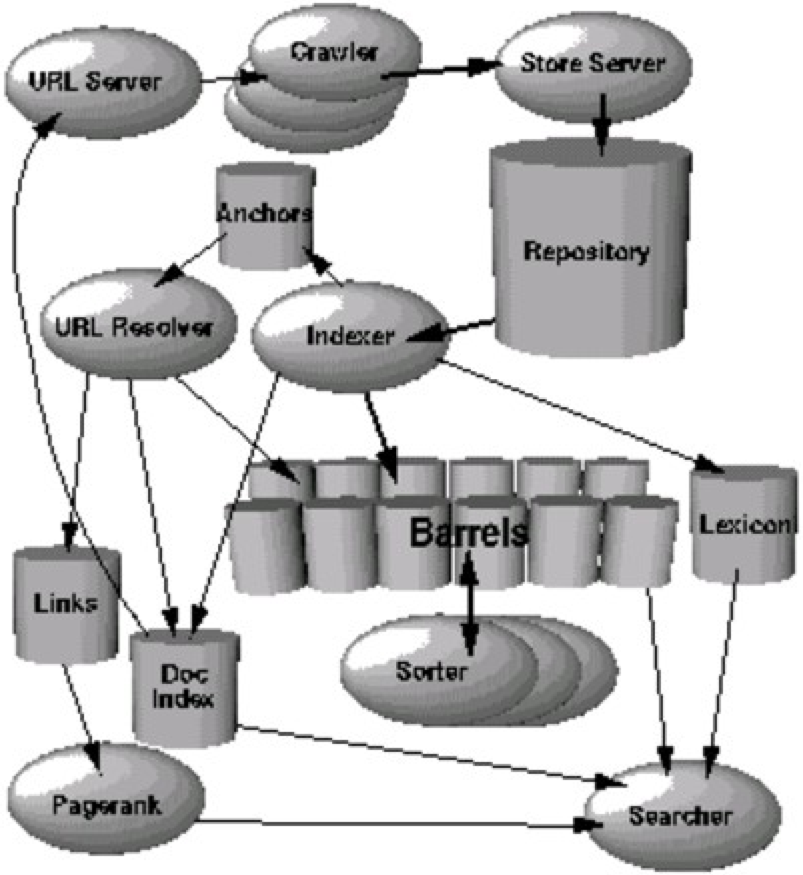
Meta-data

- Ranking of search results:
  - PageRank
  - Relevance to query
- Anchor text
  - Often describes a page better than the content itself
  - Can be abused through coordinated effort
- Other information:
  - Visual details
  - Page update frequency
  - Search term proximity

Google Architecture

- Overview
- Data Structures
- Crawling
- Indexing
- Searching

Google Architecture



Data Structures

- BigFiles
- Repository
- Document Index
- Lexicon
- Hit Lists
- Forward Index
- Inverted Index

- Crawling
- Indexing:
  - Parsing
  - Indexing into Barrels
  - Sorting

Searching

1. Parse query
2. Convert words into wordID
3. Seek start of doclist for every word
4. Scan through doclist until there is document matching all terms
5. Compute rank of document for query
6. If we are not at the end of any doclist and we haven't reached max. nr. of documents, go to 4
7. Sort matched documents by rank, return top k.

Conclusions

- Original papers present Google as research project
- Commercial success largely due to technical superiority
- Influence:
  - On everyday life
  - On businesses
- AdSense, AdWords:
  - Made Google viable commercially
  - Changed the web by providing an easy way to monetize content
- Google:
  - Monitor the web, grow with it
  - Influence the web (SEO industry)

“Do no evil”

- Privacy
- Filtering results
  - Google bombs
  - Link farms
  - Illegal stuff
  - Political issues
- Transparency (search data)
- Transparency (algorithms)

Thanks for the attention.