The Anatomy of a Large-Scale Hyper-textual Web Search Engine

Lawrance Page and Sergey Brin

Presented by
Bahar Şahin, Serhat Özcan

Outline

- Introduction
- Motivation
- Google Architecture
- Results and Performance
- Future Work
- Conclusion

Introduction

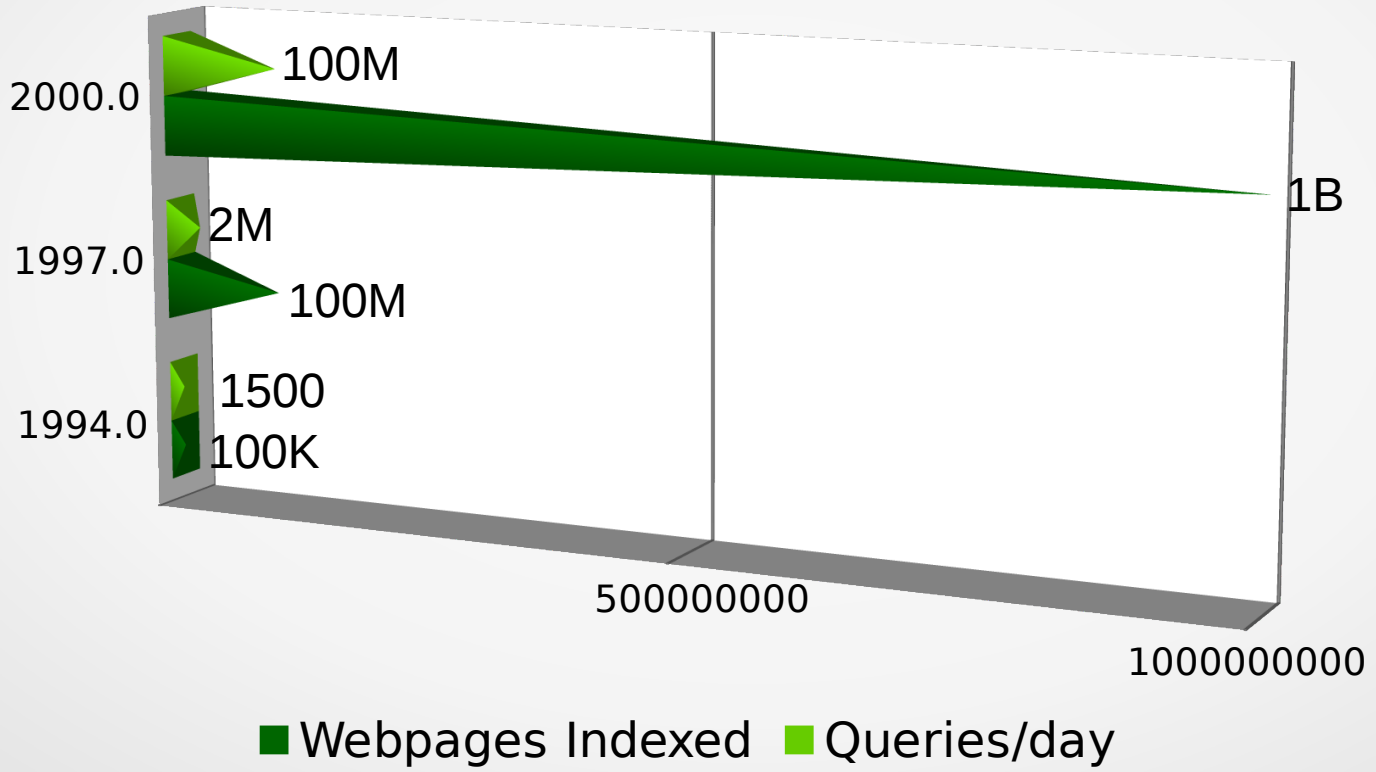
- Authors S. Brin and L. Page were PhD students at Stanford
- The original Google Paper
  - Describes a prototype search engine → Google.
- 10^{100} = Googol

Introduction: The World Wide Web

- 1994, search engine World Wide Web Worm (WWWW) indexed **110K** docs.
- 1997, Search engines claimed to index from 2M – 100M web pages.

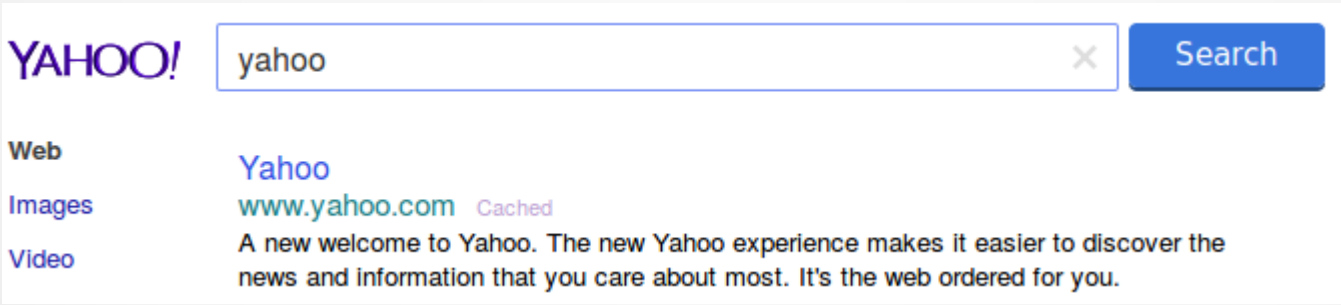
- 1997, Altavista claimed receiving **20M** queries per day.
- 1994, WWWW received average of **1500** queries per day.

Design Goal: Scaling with the Web



Design Goal : Improved Search Quality

- Search quality is very bad → junk results
- November 1997, only one of the top four commercial search engines finds itself



Design Goal : Academic Search Engine Research

- Search engine development migrated from academic domain to commercial.
- Contributing the academic development and research in search engines.
- Transparency

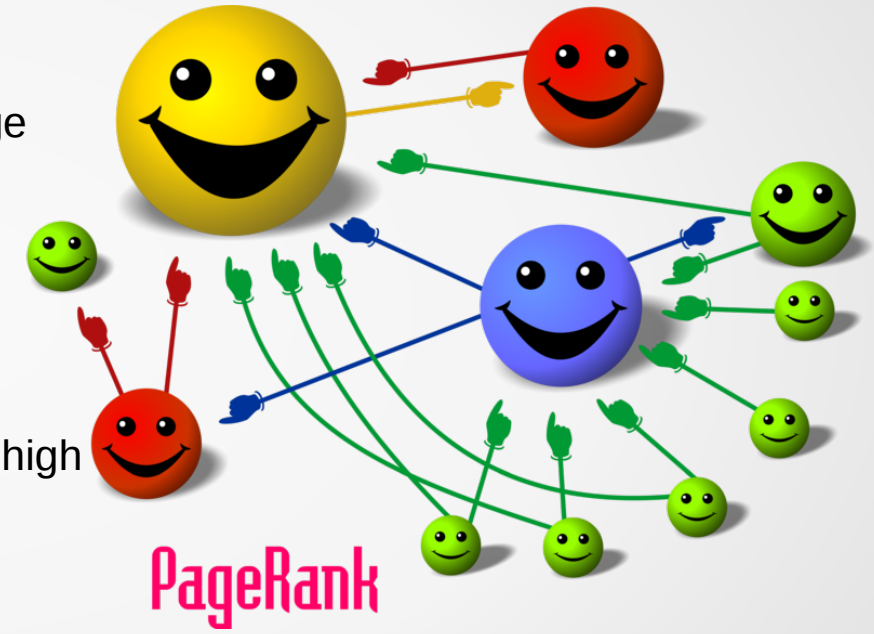
Design Goal : Build System that People Use

- Query logs are important for SE research but they are not publicly available due to its commercial value.

- Ease of use:
  - short queries → good results

System Features: PageRank

- Based on hyperlinks map
- PageRank of a webpage is:
  - probabilty of random surfer will click on the page by randomly clicking on the link
- A webpage have a high PageRank if:
  - There are many pages pointing to it
  - Or, There are some pages pointing to it having high PageRank
- Covered in another paper of Brin and Page



SystemFeatures: Anchor Text

- < a href = “http://www.bilkent.edu.tr“> “Bilkent Üniversitesi“

- Provides good description of webpages
- Anchors may exists for documents that cannot be indexed by text crawler (such as pictures, audio files)

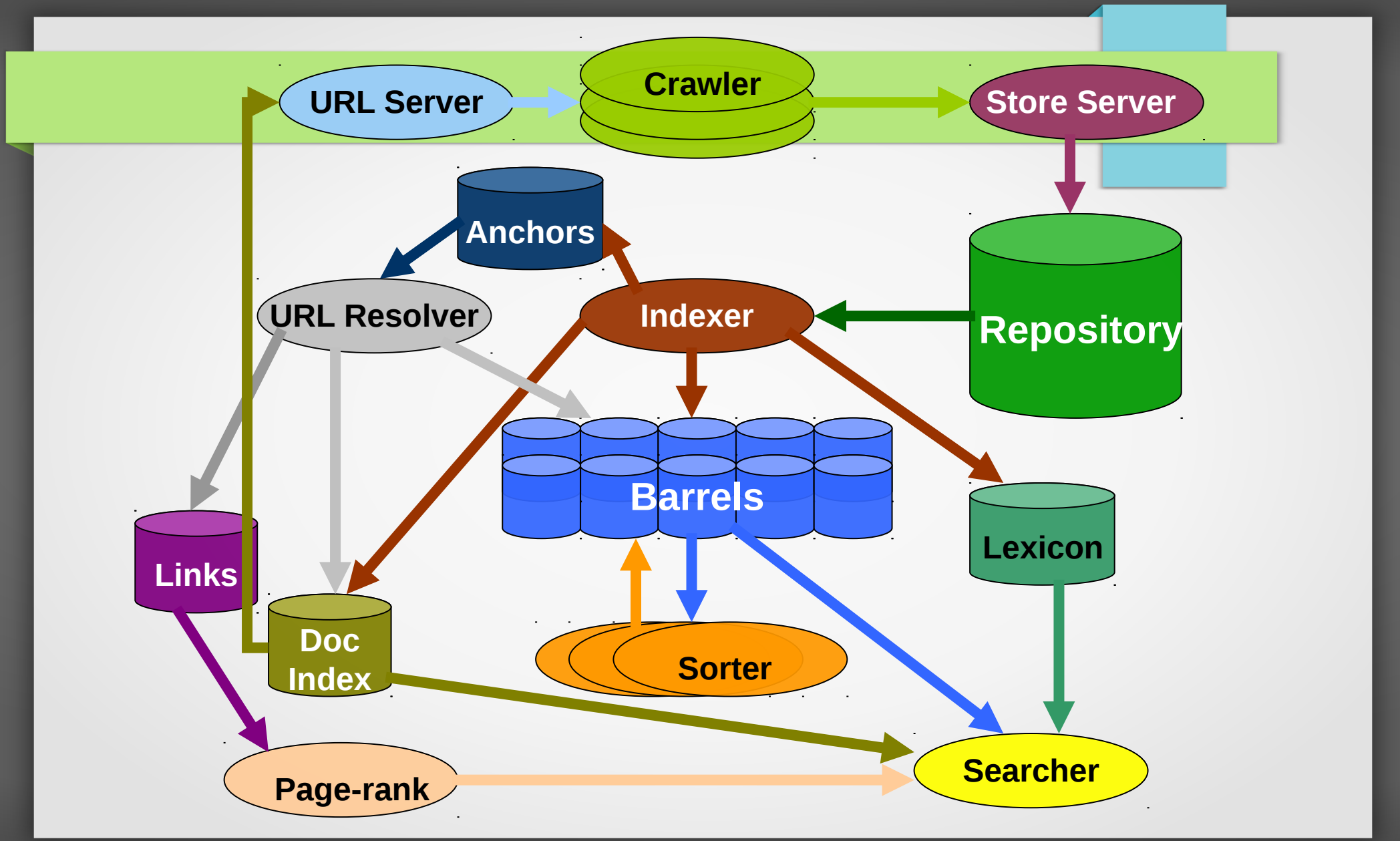
System Features: **Others**

- Word proximity in documents are taken into account.
- Word position, font size have an effect on weight of the words.

Google Architecture

Solving big problems is easier than solving little problems.

Sergey Brin


URL Server
Crawler
Store Server
Anchors
URL Resolver
Indexer
Repository
Barrels
Links
Lexicon
Doc Index
Sorter
Page-rank
Searcher

Google Architecture **Major Data Structures**

- **BigFiles**
  - Virtual file system
  - Addressable by 64b integers
- **Repository**
  - Contains the HTML
  - **zlib** library for compression
- **Lexicon**
  - List of Words (vocabulary) ~14M
  - Hash table of pointers
- **Document Index**
  - Ordered by docID and contains:
    - Document status
    - Pointer to the Repository
    - Document checksum
    - Statistics

Google Architecture Major Data Structures

- **HitList**
  - List of occurances of a word in a document
  - Contains info on position, font and capitalization.
  - Huffman Coding

- **Forward Index**
  - 64 barrels

docID	WordID =1	# Hits =2	Hit list
	WordID =2	# Hits =3	Hit list

- **Inverted Index**
  - Barrels processesed by the sorter
  - Sorted by doc id

Ranking System

1. Parse the query.
2. Convert words into wordIDs.
3. Seek to the start of the doclist in the short barrel for every word.
4. Scan through the doclists until there is a document that matches all the search terms.
5. Compute the rank of that document for the query.
6. If we are in the short barrels and at the end of any doclist, seek to the start of the doclist in the full barrel for every word and go to step 4.
7. If we are not at the end of any doclist go to step 4. Sort the documents that have matched by rank and return the top k

Results

- Authors claimed that Google is better than other search engines.
  - PageRank, anchor text, word proximity

Web Page Statistics	
Number of Web Pages Fetched	24 million
Number of Urls Seen	76.5 million
Number of Email Addresses	1.7 million
Number of 404's	1.6 million

Storage Statistics	
Total Size of Fetched Pages	147.8 GB
Compressed Repository	53.5 GB
Short Inverted Index	4.1 GB
Full Inverted Index	37.2 GB
Lexicon	293 MB
Temporary Anchor Data (not in total)	6.6 GB
Document Index Incl. Variable Width Data	9.7 GB
Links Database	3.9 GB
Total Without Repository	**55.2 GB**
Total With Repository	**108.7 GB**

Performance

- Query Evaluation → 1 to 10 seconds

	Initial Query		Same Query Repeated (IO mostly cached)	
Query	**CPU Time(s)**	**Total Time(s)**	**CPU Time(s)**	**Total Time(s)**
al gore	0.09	2.13	0.06	0.06
vice president	1.77	3.84	1.66	1.80
hard disks	0.25	4.86	0.20	0.24
search engines	1.31	9.63	1.16	1.16

Future Work

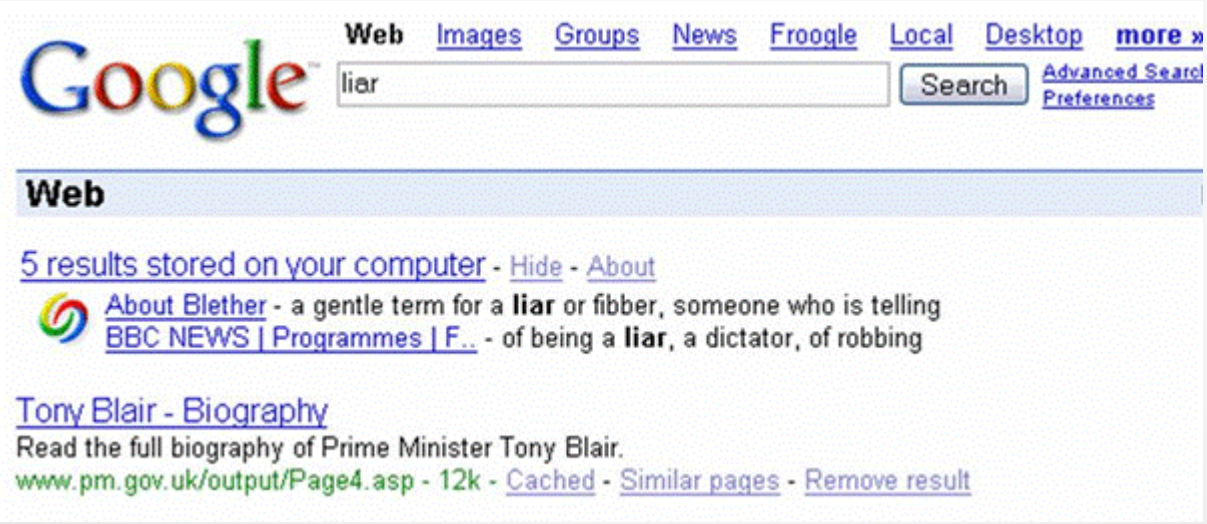
- Make Google Faster
  - Query Caching
  - Sub-indices on common terms
  - Smart disk allocation
- Improve Search Quality by implementing:
  - Clustering
  - Relevance Feedback
  - Smart Algorithms

Conclusion

Issues:

Google Bombs

"Collective efforts to link to a site by a key phrase and artificially elevate a Web site in the Google search results for that search phrase."



Conclusion

#9 Larry Page

+ Follow (1.415)

Real Time Net Worth As of 4/9/15

$29.5 Billion

#9 Sergey Brin

+ Follow (1,500)

Real Time Net Worth As of 4/9/15

$29 Billion

Cofounder, Director Of Special Projects, Google

The World's Most Powerful People