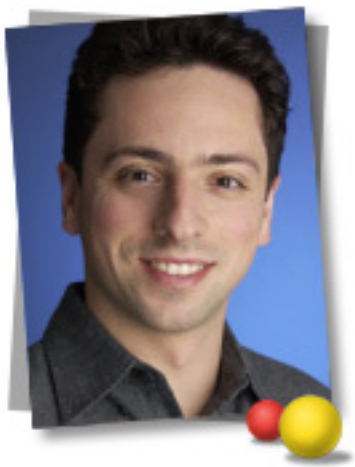
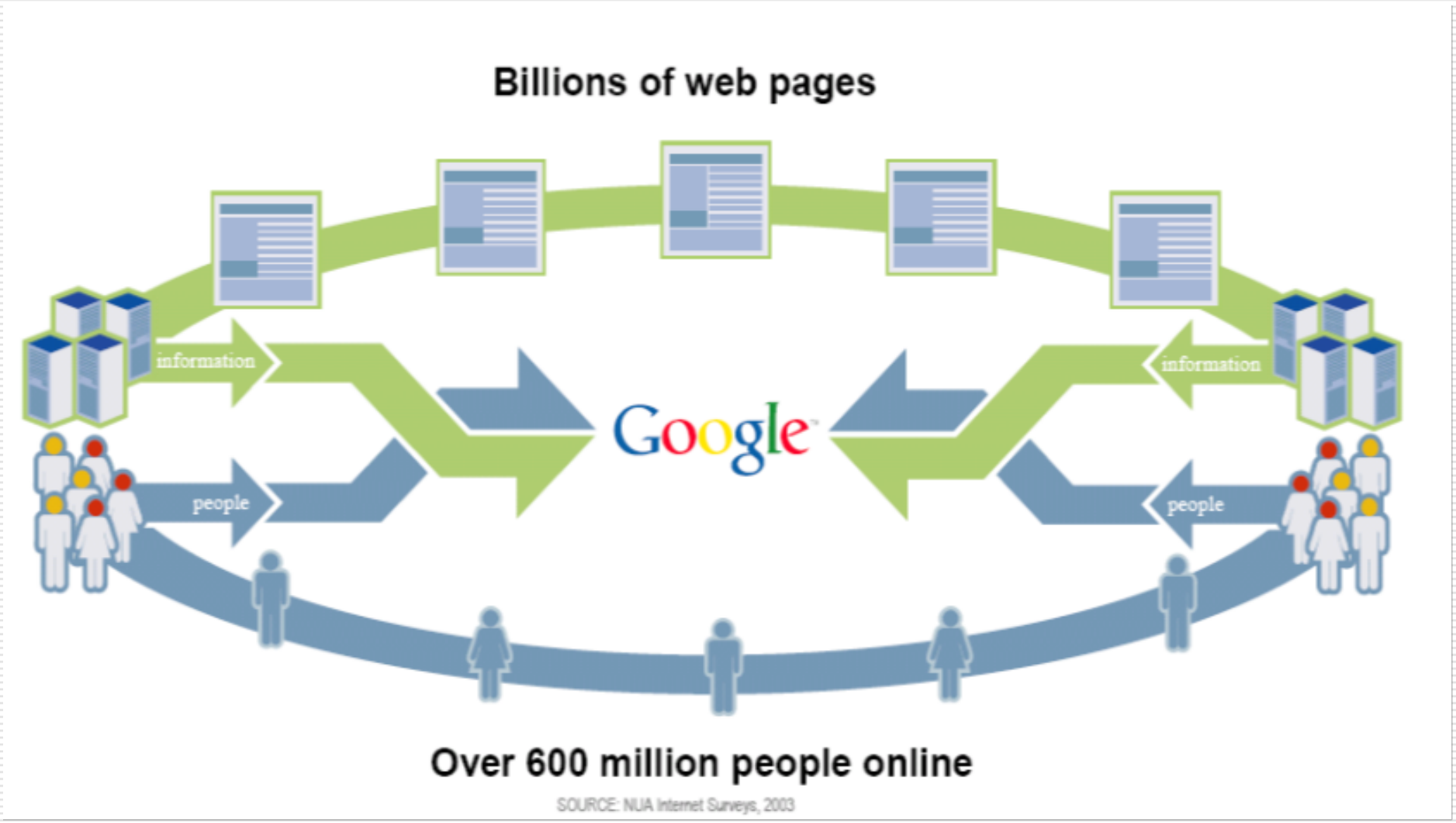
The Anatomy of a Large-Scale Hypertextual Web Search Engine



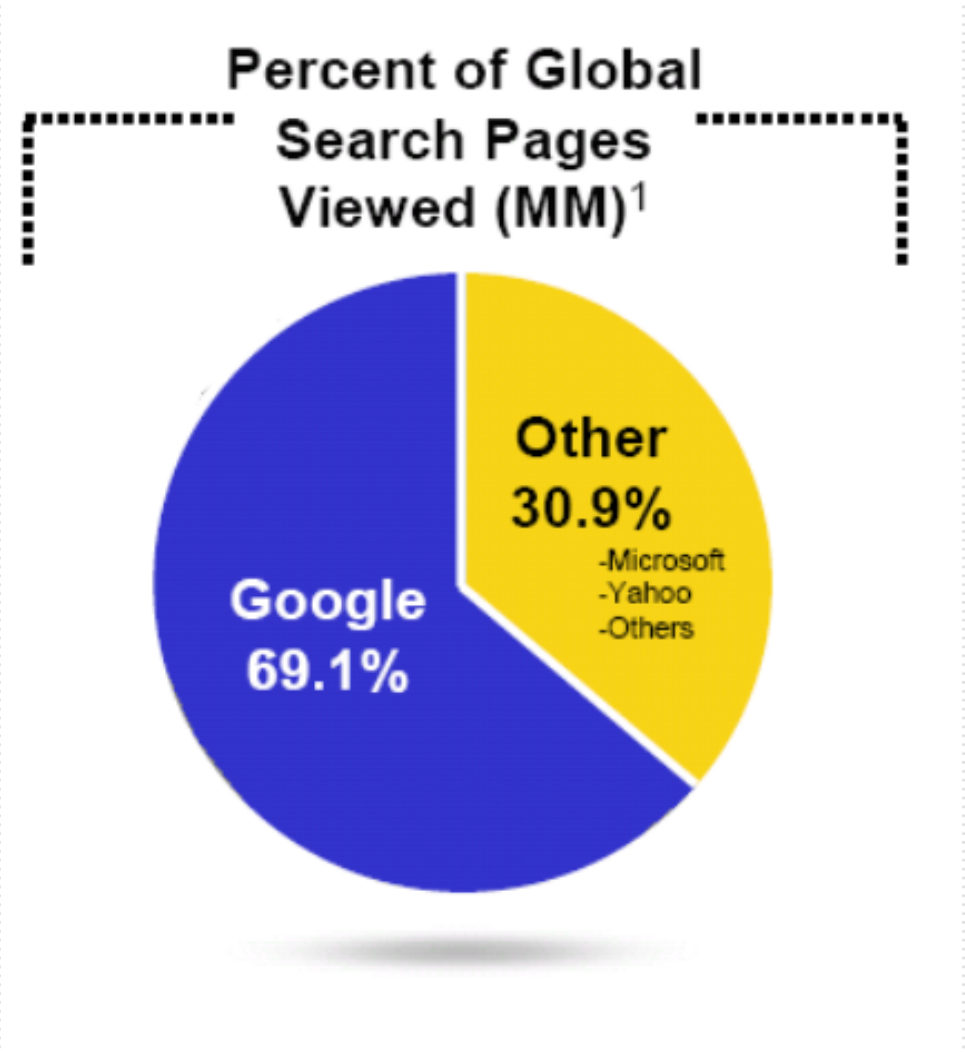
Sergey Brin
Co-Founder
& President,
Technology

Larry Page
Co-Founder &
President,
Products

Introduction



Introduction



Introduction

- Google's Mission

To organize the world's information and make it universally accessible and useful

- Scaling with the web
  - Improved Search Quality
  - Academic Search Engine Research

Google System Features

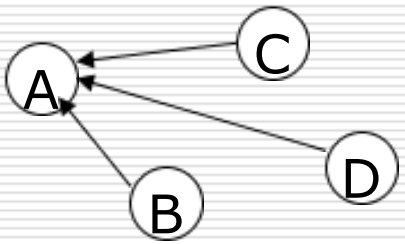
- It makes use of the link structure of the Web to calculate a quality ranking for each web page, called PageRank
- PageRank is a trademark of Google. The PageRank process has been patented.
- Google utilizes link to improve search results

PageRank

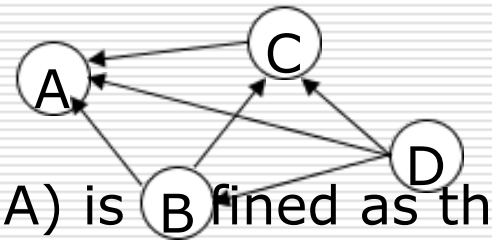
- PageRank is a link analysis algorithm which assigns a numerical weighting to each Web page, with the purpose of "measuring" relative importance.
  - Based on the hyperlinks map
  - An excellent way to prioritize the results of web keyword searches

Simplified PageRank algorithm

- Assume four web pages: **A**, **B**,**C** and **D**. Let each page would begin with an estimated PageRank of 0.25.



$$PR(A) = PR(B) + PR(C) + PR(D).$$



$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

- L(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

PageRank algorithm including damping factor

- Assume page A has pages B, C, D ..., which point to it. The parameter d is a damping factor which can be set between 0 and 1. Usually set d to 0.85. The PageRank of a page A is given as follows:

$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$

Intuitive Justification

- A "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back“, but eventually gets bored and starts on another random page.

- The probability that the random surfer visits a page is its PageRank.
- The d damping factor is the probability at each page the "random surfer" will get bored and request another random page.

- 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.

Anchor Text

- <A href="http://www.yahoo.com/">Yahoo!</A>

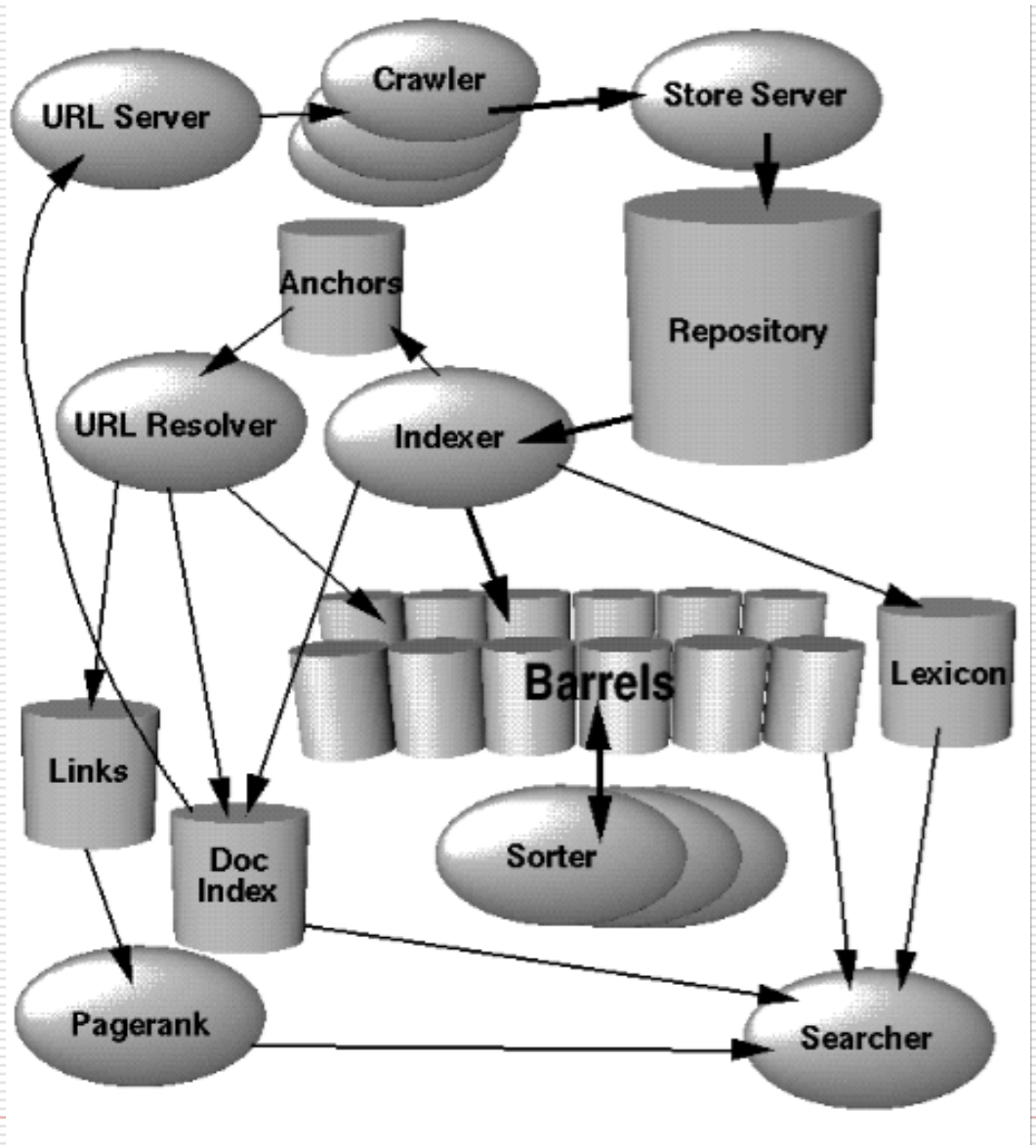
Besides the text of a hyperlink (anchor text) is associated with the page that the link is on, it is also associated with the page the link points to.

- anchors often provide more accurate descriptions of web pages than the pages themselves.
- anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases.

Other Features

- It has location information for all hits.
- Google keeps track of some visual presentation details such as font size of words.
  - Words in a larger or bolder font are weighted higher than other words.
- Full raw HTML of pages is available in a repository

Google Architecture Overview



Major Data Structures

- **BigFiles**
  - virtual files spanning multiple file systems and are addressable by 64 bit integers.
- **Repository**
  - contains the full HTML of every web page.
- **Document Index**
  - keeps information about each document.
- **Lexicon**
  - two parts – a list of the words and a hash table of pointers.
- **Hit Lists**
  - a list of occurrences of a particular word in a particular document including position, font, and capitalization information.
- **Forward Index**
  - **stored in a number of barrels**
- **Inverted Index**
  - consists of the same barrels as the forward index, except that they have been processed by the sorter.

Crawling the Web

- Google has a fast distributed crawling system.
- A single URLserver serves lists of URLs to a number of crawlers.
- Both the URLserver and the crawlers are implemented in Python.
  - Each crawler keeps roughly 300 connections open at once. At peak speeds, the system can crawl over 100 web pages per second using four crawlers. This amounts to roughly 600K per second of data.
- Each crawler maintains a its own DNS cache so it does not need to do a DNS lookup before crawling each document.

Indexing the Web

- Parsing
  - Any parser which is designed to run on the entire Web must handle a huge array of possible errors.
- Indexing Documents into Barrels
  - After each document is parsed, it is encoded into a number of barrels. Every word is converted into a wordID by using an in-memory hash table -- the lexicon.
  - Once the words are converted into wordID's, their occurrences in the current document are translated into hit lists and are written into the forward barrels.
- Sorting
  - the sorter takes each of the forward barrels and sorts it by wordID to produce an inverted barrel for title and anchor hits and a full text inverted barrel.

Google Searching

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.

Figure 4. Google Query Evaluation

Results and Performance

- The current version of Google answers most queries in between 1 and 10 seconds.
- The table shows some samples search time from the current version of Google. They are repeated to show the speedups resulting from cached IO.

	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

Conclusion

- Google is designed to be a scalable search engine.
- The primary goal is to provide high quality search results over a rapidly growing World Wide Web.
- Google employs a number of techniques to improve search quality including page rank, anchor text, and proximity information.
- Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them.

Google bomb

- Because of the PageRank, a page will be ranked higher if the sites that link to that page use consistent anchor text.
- A Google bomb is created if a large number of sites link to the page in this manner.
- search term "more evil than Satan himself"
  - the Microsoft homepage as the top result.

Problems

- High Quality Search
  - The biggest problem facing users of web search engines today is the quality of the results they get back.
- Scalable Architecture
  - Google is designed to scale. It must be efficient in both space and time

The Future

"The ultimate search engine would understand exactly what you mean and give back exactly what you want."

- Larry Page

Thanks
!