Outline

- Historical Background
- Information Tsunami
- Anatomy of a Web Page
- Anatomy of Web Access
- The Challenge of Search
- Google's Page Rank Algorithm
- Fun and Games with Internet Search
- New Directions

Needle in the Haystack:
The Technology of Internet Search

Randy H. Katz
The United Microelectronics Corporation Distinguished Professor
Computer Science Division, EECS Department
University of California, Berkeley
Berkeley, CA 94720-1776 USA
randy@cs.berkeley.edu

Search is BIG!

And the World is Going Digital

Historical Background: The Perfect Storm

World Wide Web

Arpanet 1969
NFS 1983
Gopher 1992
WWW 1990
Netscape 1993
Explorer 1995
PageRank 1998
YAHOO 1995
THOMAS 1996
Iphone 2007

Est. $15.5 Billion spent online
Thanksgiving to Xmas 2004
up 28% since 2003
Outline

• Historical Background
• Information Tsunami
• Anatomy of a Web Page
• Anatomy of Web Access
• The Challenge of Search
• Google’s Page Rank Algorithm
• Fun and Games with Internet Search
• New Directions

Information Tsunami

• Bit: Binary digit – either a 0 or 1
• Byte: 8 bits
  – 1 byte: single character
  – 10 bytes: a single word
  – 100 bytes: a single typed page
• Kilobyte: 1,000 or 10^3 bytes
  – 1 kilobyte: Very short story
  – 2 kilobytes: Typewriter page
  – 50 kilobytes: Complete works of Shakespeare
• Megabyte: 1,000,000 or 10^6 bytes
  – 1 megabyte: Typical work of Shakespeare
  – 2 megabytes: By publishing year
  – 10 megabytes: Complete works of Shakespeare
• Gigabyte: 1,000,000,000 or 10^9 bytes
  – 1 gigabyte: The United States
  – 2 gigabytes: Hi-res photo
  – 5 gigabytes: Encyclopedias
• Terabyte: 1,000,000,000,000 or 10^12 bytes
  – 1 terabyte: A typical year’s worth of information
  – 2 terabytes: Academic research library
• Exabyte: 1,000,000,000,000,000,000 or 10^18 bytes
  – 2 exabytes: Total human knowledge
• Zettabyte: 1,000,000,000,000,000,000,000 or 10^21 bytes
• Yottabyte: 1,000,000,000,000,000,000,000,000 or 10^24 bytes

Information Tsunami

• Megabyte: 1,000,000 or 10^6 bytes
  – 1 megabyte: Small movie or 3.5” floppy disk
  – 2 megabytes: CD
• Gigabyte: 1,000,000,000 or 10^9 bytes
  – 1 gigabyte: Hard disk
• Terabyte: 1,000,000,000,000 or 10^12 bytes
  – 1 terabyte: Magnetic tape
• Exabyte: 1,000,000,000,000,000,000 or 10^18 bytes
  – 2 exabytes: Total mass of the observable universe
• Zettabyte: 1,000,000,000,000,000,000,000 or 10^21 bytes
  – 2 zettabytes: Total mass of the observable universe

Information Tsunami

• Megabyte: 1,000,000 or 10^6 bytes
  – 1 megabyte: Small movie or 3.5” floppy disk
  – 2 megabytes: CD
• Gigabyte: 1,000,000,000 or 10^9 bytes
  – 1 gigabyte: Hard disk
• Terabyte: 1,000,000,000,000 or 10^12 bytes
  – 1 terabyte: Magnetic tape
• Exabyte: 1,000,000,000,000,000,000,000 or 10^18 bytes
  – 2 exabytes: Total mass of the observable universe
• Zettabyte: 1,000,000,000,000,000,000,000,000 or 10^21 bytes
  – 2 zettabytes: Total mass of the observable universe
• Yottabyte: 1,000,000,000,000,000,000,000,000,000 or 10^24 bytes

Outline

• Historical Background
• Information Tsunami
• Anatomy of a Web Page
• Anatomy of Web Access
• The Challenge of Search
• Google’s Page Rank Algorithm
• Fun and Games with Internet Search
• New Directions

Anatomy of a Web Page:
Randy's Home Page

URL: Uniform Resource Locator
Images
Text

Anatomy of a Web Page:
Randy's Web Page

Outline
- Historical Background
- Information Tsunami
- Anatomy of a Web Page
- Anatomy of Web Access
- The Challenge of Search
- Google's Page Rank Algorithm
- Fun and Games with Internet Search
- New Directions

Anatomy of Web Access

Web Page
In HTML

Naming System (DNS):
Name-To-Address Mapping
IP address

Token

Web Browser
http://www.umc.com.tw/

Web Server

Taiwan
Anatomy of Web Access

### Outline
- Historical Background
- Information Tsunami
- Anatomy of a Web Page
- Anatomy of Web Access
- The Challenge of Search
- Google's Page Rank Algorithm
- Fun and Games with Internet Search
- New Directions

### Challenges of Search
- How to find all the pages on the Web?
- How to order the pages by relevance?
- How to make searchable the content on those pages?
- How to keep it all up-to-date?

- **Web Crawlers/Spiders**
  - Follow links in the Web to find content
  - Web pages "scraped" for more links to follow
  - Web re-visited on the order of once every two-three days

- **Indexes**
  - Web pages "searched" for search terms to build indexes
  - Google Page ranks a page based on its position in the index, which is used to determine its relevance

### Search Challenges and Issues
- Web growing faster than search engines can index
- Web pages updated frequently, forcing frequent re-indexing
- Keyword only searches result in many false positives
- Difficulty indexing dynamically generated sites: the so-called "invisible web"
- Some search engines order results by financial "placement" considerations rather than relevance
- Some sites trick search engines to display them first for some keywords—results in polluted search results, with more relevant links pushed down among the results
### Page Ranking Algorithms

- **Web page relevancy**
  - Many hits, how to ensure the best/most relevant web page are presented first in answer to a search

- **Location and Frequency of Keywords**
  - Index terms in page title make its relevance for the term
  - Keywords near “top” of page more relevant than bottom
  - High keyword frequency boosts relevance

- If search engine strategy is known, page developers will “game” the strategy to get their pages ranked higher

### Google’s Page Rank Algorithm

- **Basic idea:**
  - PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page’s value. Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves “important” weigh more heavily and help to make other pages “important.”

- Formula:
  - \[ PR(P) = (1 – d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right) \]

- See [http://www-db.stanford.edu/~backrub/google.html](http://www-db.stanford.edu/~backrub/google.html)

### Google Server Architecture

- Index servers: search term partitioned and mapped to doc list
- Intersect to find document list: sort by page rank
- Document IDs used to extract text from Doc Servers
- Over 100,000 processors (and growing) in Googleplex
Outline

• Historical Background
• Information Tsunami
• Anatomy of a Web Page
• Anatomy of Web Access
• The Challenge of Search
• Google's Page Rank Algorithm
• Fun and Games with Internet Search
• New Directions

Fun and Games

• Google Scholar
• Googling Someone
• Google News
• Comparison Shopping
• Google Whacks
Comparison Shopping
elgooG
Google Whacks
Business Model
Ad Placement and Click-Thru

Outline
• Historical Background
• Information Tsunami
• Anatomy of a Web Page
• Anatomy of Web Access
• The Challenge of Search
• Google's Page Rank Algorithm
• Fun and Games with Internet Search
• New Directions

Top 10 Search Engines
10. DMOZ.org
9. AlltheWeb.com
8. KariOIO.com
7. MSN.com
6. Dogpile.com
5. AskJeeves.com
4. About.com
3. Yahoo.com
2. Vivisimo.com
1. Google.com

Old data (2002): Google is now market leader in ad revenue
2004 revenue through 9/30/04: $2.1B
Innovations Now and Yet to Come

- Index even larger portions of the Web, even beyond traditional web pages, e.g., video
- Better quality/higher relevance searches
- Better presentation of results, e.g., clustering, site information
- Better exploitation of semantic relationships for improved page ranking, more personalization, e.g., user's zip code
- More services (Web, news groups, blogs, comparison shopping, video/audio, yellow pages, etc.)
- Integrate with desktop machine

Parting Thoughts

"Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"


"Needle in the Haystack: The Technology of Internet Search"