CS155b: E-Commerce

Lecture 15: March 6, 2003

Web Searching and Google

Finding Information on the Internet

The Internet is so successful partly because it is so easy to publish information on the World Wide Web.

- No central authority on what pages exist, where they exist, or when they exist.
- Too much to sort through, anyway.
- **Question: How do we find what we need on the web?**

WWW Search Engines

- **Answer**: Set up websites that people can use to search for information by performing a ***search query***.
- Not such an easy solution! In addition to the technical problems, we have these business questions:
  - How do people know about the search engine websites?
  - How do you make money off of this? (Especially now that the service is free.)

Examples of Search Websites

- Website directories that have grown to become portals
  - Yahoo! (first searches its own hand-made directory, then Google index)
  - Lycos
  - Excite
- ISP + portals that now include search
  - AOL / Netscape (agreement with Google, as of 6/2002)
  - MSN (agreement with Inktomi – the search engine technology also used by Yale's website)
- InfoSpace / MetaCrawler, a "search engine searcher"
- AskJeeves, a "natural language" search engine
- Google, a "traditional search" website that remains dedicated to searching

Solutions (?) to Technical Problems

- How do we keep track of what pages are on the WWW?
  - Have a *crawler* or *spider* scan the web and links between pages to find new, updated, and removed pages.
- How do we store the content we find?
  - Design a way to map keywords in queries to documents so we can return a *usefully ordered list* to the user.
- What happens when pages are temporarily unavailable?
  - Use *caching*: keep a local copy of documents as we crawl the web. (Need lots of space!)

Solutions (?) to Technical Problems *(continued)*

- How do we store all the information?
  - Use a large network of disks (and maybe a clever method of compression) that can be easily searched.
- How do we handle so many different requests?
  - Use a *cluster* of computers that work together to process queries.

There is still ongoing research to find better ways to solve these problems!

WWW Digraph

- More than 3 Billion Nodes (Pages)
- Average Degree (links/Page) is 5-15. (Hard to Compute!)
- Massive, *Distributed*, *Explicit* Digraph (Not Like Call Graphs)

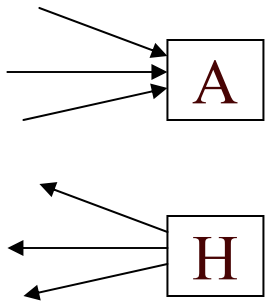
"Hot" Research Area

- Graph Representation
- Duplicate Elimination
- Clustering
- Ranking Query Results

"Abundance" Problem

http://simon.cs.cornell.edu/home/kleinber/kleinber.html

- Given a query find:
  - Good Content ("Authorities")
  - Good Sources of Links ("Hubs")
- Mutually Reinforcing
- Simple (Core) Algorithm



$T \hat{=} \{n \text{ Pages}\}$, $A \hat{=} \{\text{Links}\}$

$X_p \varepsilon \Re \geq 0$, $p \varepsilon T$ non-negative "Authority Weights"

$Y_p \varepsilon \Re \geq 0$, $p \varepsilon T$ non-negative "Hub Weights"

I operation Update Authority Weights

$$X_p \leftarrow \sum_{(q,p) \varepsilon A} Y_q$$

O operation Update Hub Weights

$$Y_p \leftarrow \sum_{(p,q) \varepsilon A} X_q$$

Normalize: $\sum_{p \varepsilon T} X_p^2 = \sum_{p \varepsilon T} Y_p^2 = 1$

Core Algorithm

$Z \leftarrow (1,1,\ldots,1)$

$X \leftarrow Y \leftarrow Z$

Repeat until Convergence

  Apply I /* Update Authority weights */

  Apply O /* Update Hub Weights */

  Normalize

Return Limit (X^*, Y^*)

Convergence of

$(X^i, Y^i) \triangleq (OI)^i(Z,Z)$

$A \triangleq$ n x n "Adjacency Matrix"

Rewrite I and O:

$$X \leftarrow A^T Y \quad ; \quad Y \leftarrow AX$$

$$X^i = (A^T A)^{i-1} A^T Z \quad ; \quad Y^i = (AA^T)^i Z$$

AA^T Symm., Non-negative and $Z = (1,1,\ldots,1) \Rightarrow$

$$X^* \triangleq \lim_{i \to \infty} X^i = \omega_1(A^T A)$$

$$Y^* \triangleq \lim_{i \to \infty} Y^i = \omega_1(AA^T)$$

Whole Algorithm (k,d,c)

$q \Rightarrow$ [Search Engine] $\Rightarrow |S| \leq k$

Base Set T:

(In S, S → , → S) and ≤ d links/page

Remove "Internal Links"

Run Core Algorithm on T

From Result (X,Y), Select

C pages with max X* values

C pages with max Y* values

Examples (k= 200, d=5)

q = censorship + net

- www.EFF.org
- www.EFF.org/BlueRib.html
- www.CDT.org
- www.VTW.org
- www.ACLU.prg

q = Gates

- www.roadahead.com
- www.microsoft.com
- www.ms.com/corpinfo/bill-g.html

[Compares well with Yahoo!, Galaxy, etc.]

Approach to "Massiveness": Throw Out Most of G!!

- Non-principal Eigenvectors correspond to "Non-principal Communities"
- Open (?):

  Objective Performance Criteria

  Dependence on Search Engine

  Nondeterministic Choice of S and T

- Full name: Google, Inc.
- Privately held company. Funding partners include Kleiner Perkins Caufield & Byers and Sequoia Capital.
- Employees: over 500 worldwide (more than 50 with Ph.D.)
- Mission: "[To] deliver the best search experience on the Internet by making the world's information universally accessible and useful."
- Award-winning search engine that has indexed over 3 billion web pages (note: index size 1.6B in 12/2001.)

Google History

- 1998: Founders Larry Page and Sergey Brin (Ph.D. students at Stanford) raise $1 million from family, friends, and angel investors. Google is incorporated Sept. 7. Site receives 10,000 queries per day and is listed in PC Magazine's top 100 search websites list.
- 1st half 1999: Google has 8 employees and answers 500,000 queries/day. Red Hat (Linux distributor) becomes first customer. Google gets $25 million equity funding.

Google History *(continued)*

- 2nd half 1999: 39 employees, 3 million queries/day. Partners with Virgilio of Italy to provide search services.

- 2000: Becomes largest web search engine, having indexed 1 billion documents. Answers 18 million queries/day. Gains more partners, including Yahoo! Starts web directory.

Google History *(continued)*

- 2001: Acquires Deja.com's Usenet archive, adding newsgroups to Google's index.
  Improves and adds services including browser plug-ins, image searching, PDF searching, cell-phone and handheld compatibility, and queries and document searches in many languages.
  Advertising services used by over 350 Premium Sponsorship customers.

- Current: 3 billion web pages, 22 million PDF files, 700 million newsgroup messages, and 425 million images indexed.
  Serves 150 million queries/day.

Google Partners

- Yahoo!
- Palm
- Nextel
- Netscape
- Cisco Systems
- Virgin Net
- Netease.com
- RedHat
- Virgilio
- Washingtonpost.com

Google's Business Model

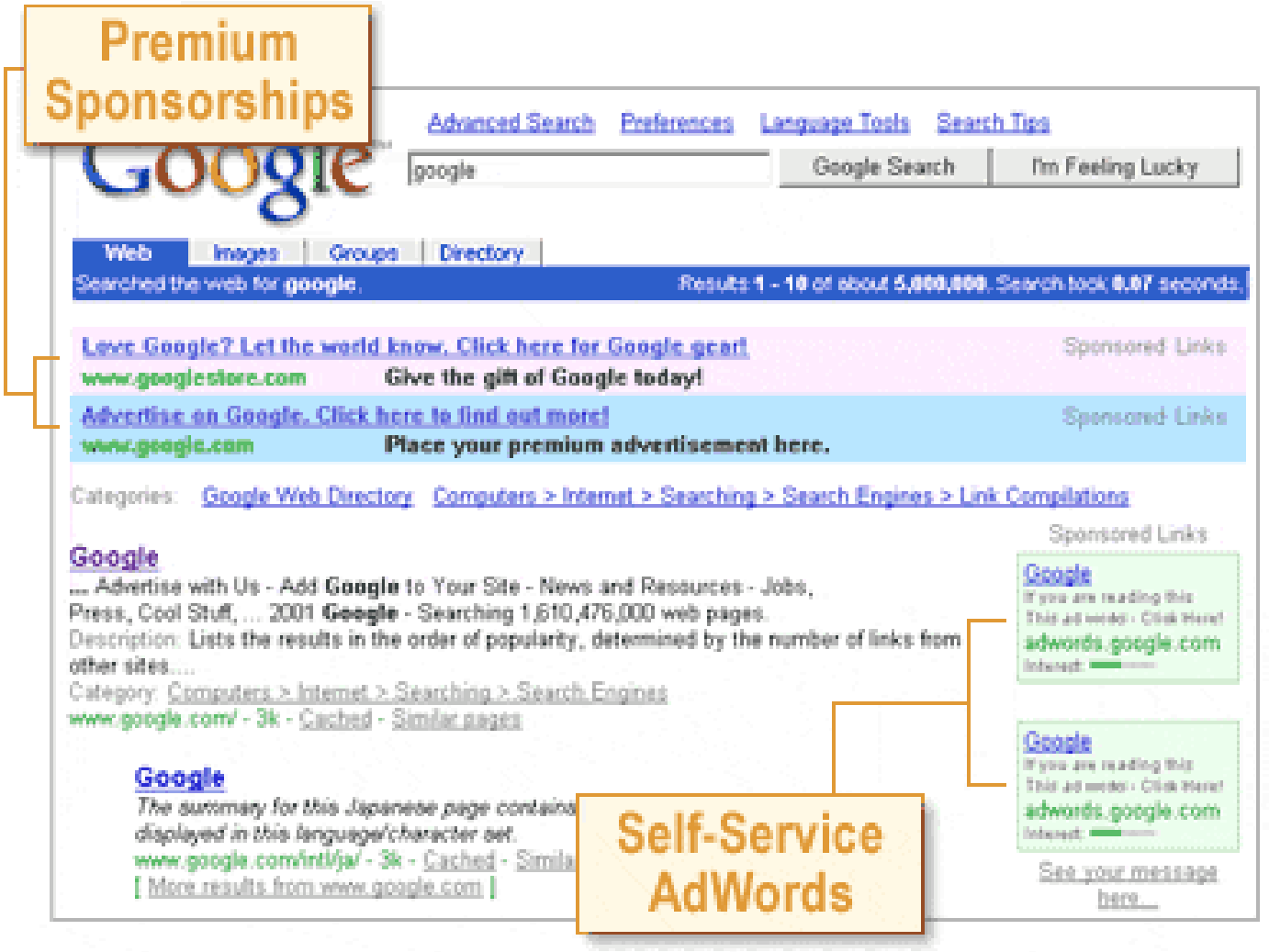
Scalable Search Services:

- Google provides customized search services for websites.
- Has become the primary search engine used by popular portal and ISP websites.

Advertising:

- *Premium Sponsorship:* sponsored text links at the top of search results based on search category.
- *AdWords:* keyword-targeted, self-service advertising method. Choose keywords or phrases where text ads will appear to the right of the search result list.
- No banner ads or graphics!

Google Advertising Screenshot



Technical Highlights

- **PageRank Technology:** Heavily mathematical (linear algebra!), objective calculation of the *PageRank* (=importance?) of a page.
  - A link from Page A to Page B is a "vote" for B.
  - The importance of A is factored into the vote.
  * Unlike other search engines, businesses cannot pay to modify PageRank results. (Note that employees can, sometimes, but only in special cases like hiding sensitive data by special request.)
- **Hypertext-Matching Analysis:** The HTML tags are taken into account when examining the contents of a page. Headings, fonts, positions, and content of neighboring pages influence the analysis.

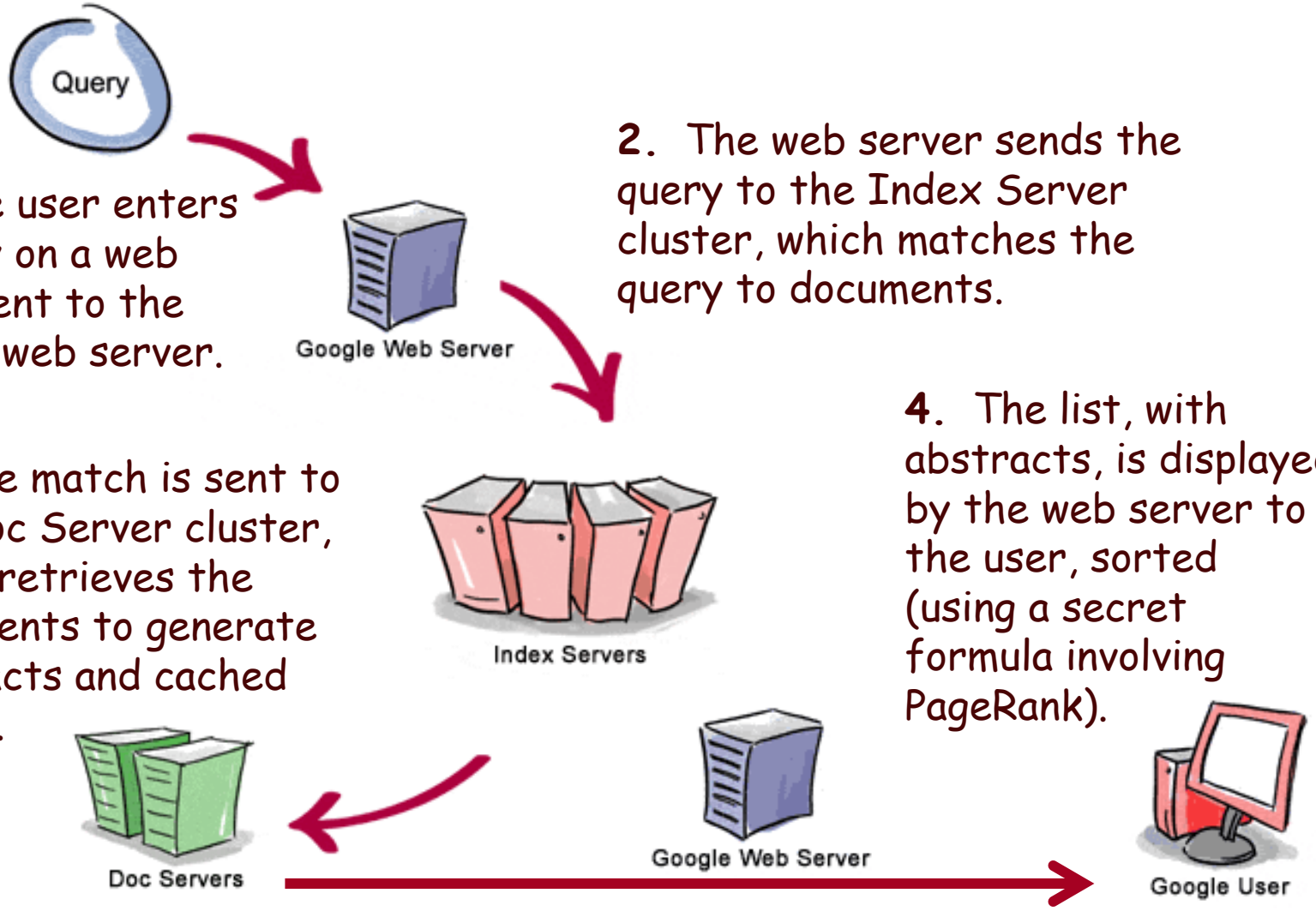
Tech Highlights *(continued)*

- **Scalable Core Technology:** Calculations are performed by the largest commercial Linux cluster of over 10,000 servers. (See the new edition of the Hennessy & Patterson computer architecture textbook for more information.) *Can grow with the Internet!*
- **Complex-File Searching:** Google can now index files in "non-Internet" formats, *e.g.*:
  - PostScript, PDF (Adobe)
  - Word, Excel, PowerPoint, Works (Microsoft)
  - WordPro, 1-2-3 (IBM/Lotus SmartSuite)
  - MacWrite
  - Rich Text (RTF), plain text

Tech Highlights *(continued)*

- **Bayesian Spelling-Suggestion Program:** Offers suggestions for misspelled words in queries, making searching easier. ("*Did you mean…?*")
- **Internationalization:**
  - Google is developing technology to index pages with complex scripts, *e.g.*:
    - Some East Asian languages have no spaces between words.
    - Hebrew and Arabic are written right-to-left; Chinese is sometimes top-to-bottom.
  - Google has a translation engine and provides its interface in many languages.
  - Current research question: How to detect the language(s) of a page?

Life of a Query

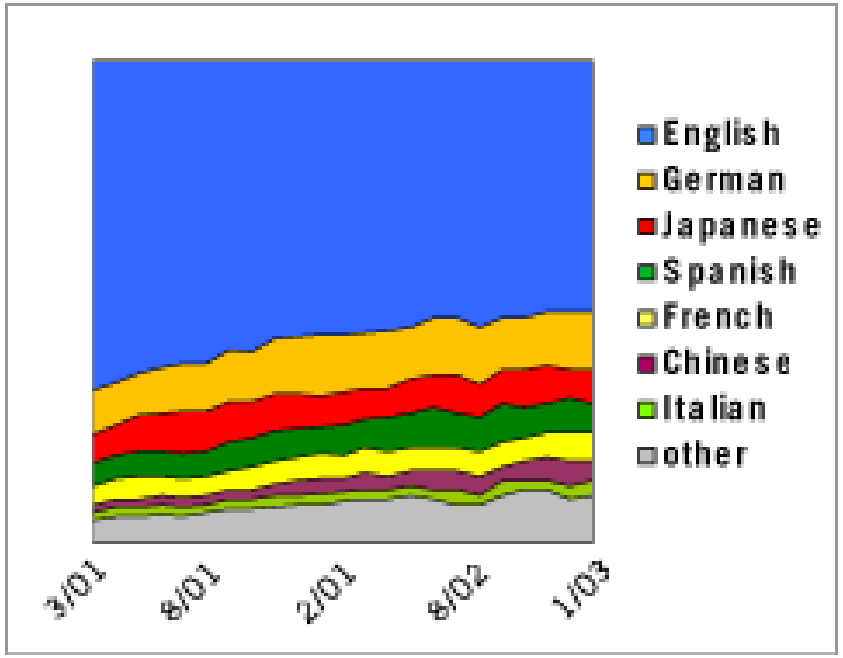


Searching Habits

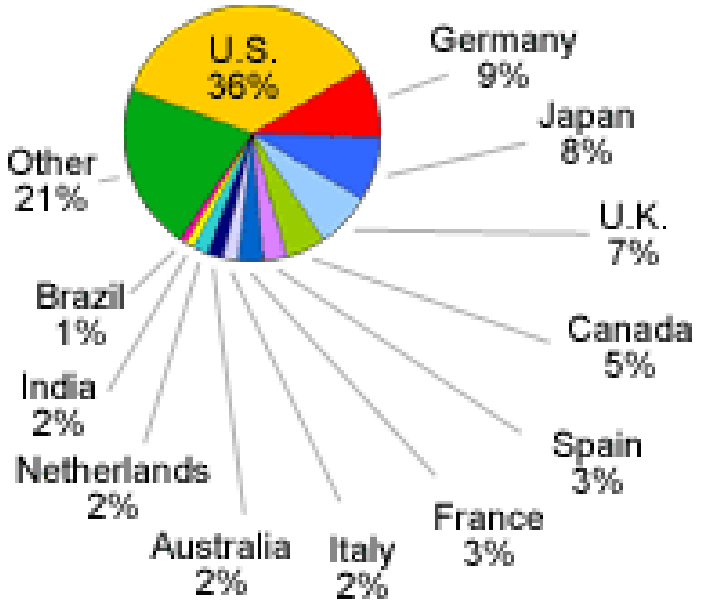
Google's *Zeitgeist* has interesting statistics about people's searches by logging the search queries!

http://www.google.com/press/zeitgeist.html

Languages used to search Google (March 2001 – January 2003)



Origin of Google searches by country (October 2001)



Searching Habits *(continued)*

Top Ten Gaining Queries (Week Ending 2/25/03)

1. great white
2. grammys
3. bachelorette
4. norah jones
5. mike tyson
6. john mayer
7. sports illustrated
8. egunkaria
9. brit awards
10. earthquake

Top Ten Declining Queries (Week Ending 2/25/03)

1. valentines day
2. joe millionaire
3. frenchie davis
4. westminster dog show
5. weather channel
6. flowers
7. 3dmark 2003
8. cricket world cup
9. curt hennig
10. jennifer garner

Top Ten Brand Names Searched: (Year, 2002)

1. Ferrari
2. Sony
3. Nokia
4. Disney
5. Ikea
6. Dell
7. Ryanair
8. Microsoft
9. Porsche
10. HP