# CS155a: E-Commerce

Lecture 20: November 27, 2001

## Web Searching and Google

# Finding Information on the Internet

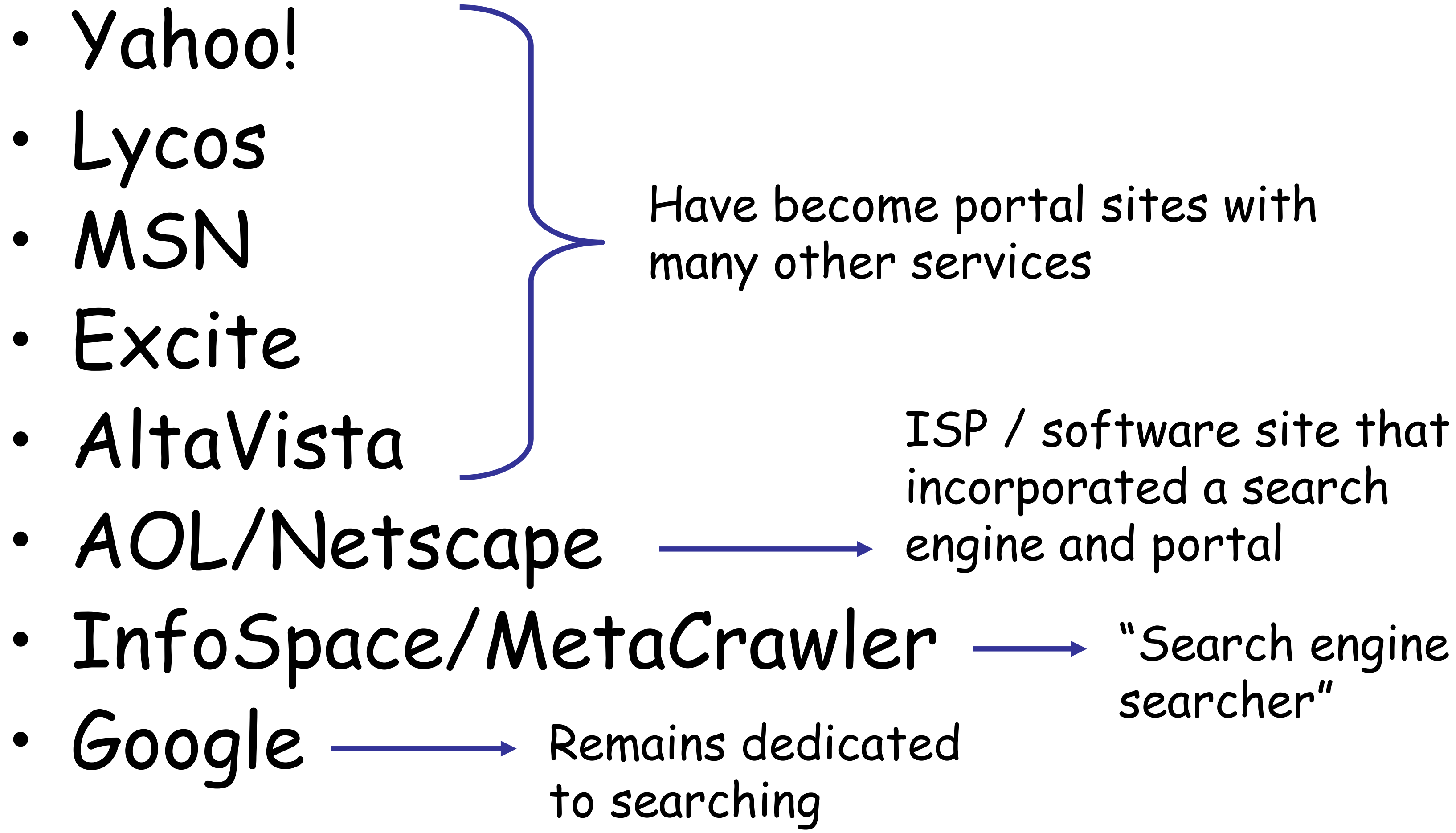
The Internet is so successful partly because it is so easy to publish information on the World Wide Web.

- 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 Engines



- Yahoo!
- Lycos
- MSN
- Excite
- AltaVista
- AOL/Netscape
- InfoSpace/MetaCrawler
- Google

Have become portal sites with many other services

ISP / software site that incorporated a search engine and portal

"Search engine searcher"

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

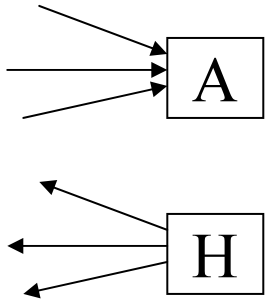
# “Hot” Research Area

- 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



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

$X_p \;\varepsilon\; \Re^{\geq} 0,\; p \;\varepsilon\; T$ 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: $\displaystyle \sum_{p\;\varepsilon\;T} X_p^2 = \sum_{p\;\varepsilon\;T} Y_p^2 = 1$

# 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:

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

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

$AA^T$ 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)

$q \Rightarrow \boxed{\text{Search Engine}} \Rightarrow |S| \leq k$

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

# Google™

- Full name: Google, Inc.
- Privately held company. Funding partners include Kleiner Perkins Caufield & Byers and Sequoia Capital.
- Employees: over 260
  (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 1.6 billion web pages.

# 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: 1.6 billion web pages, 22 million PDF files, 650 million newsgroup messages, and 250 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

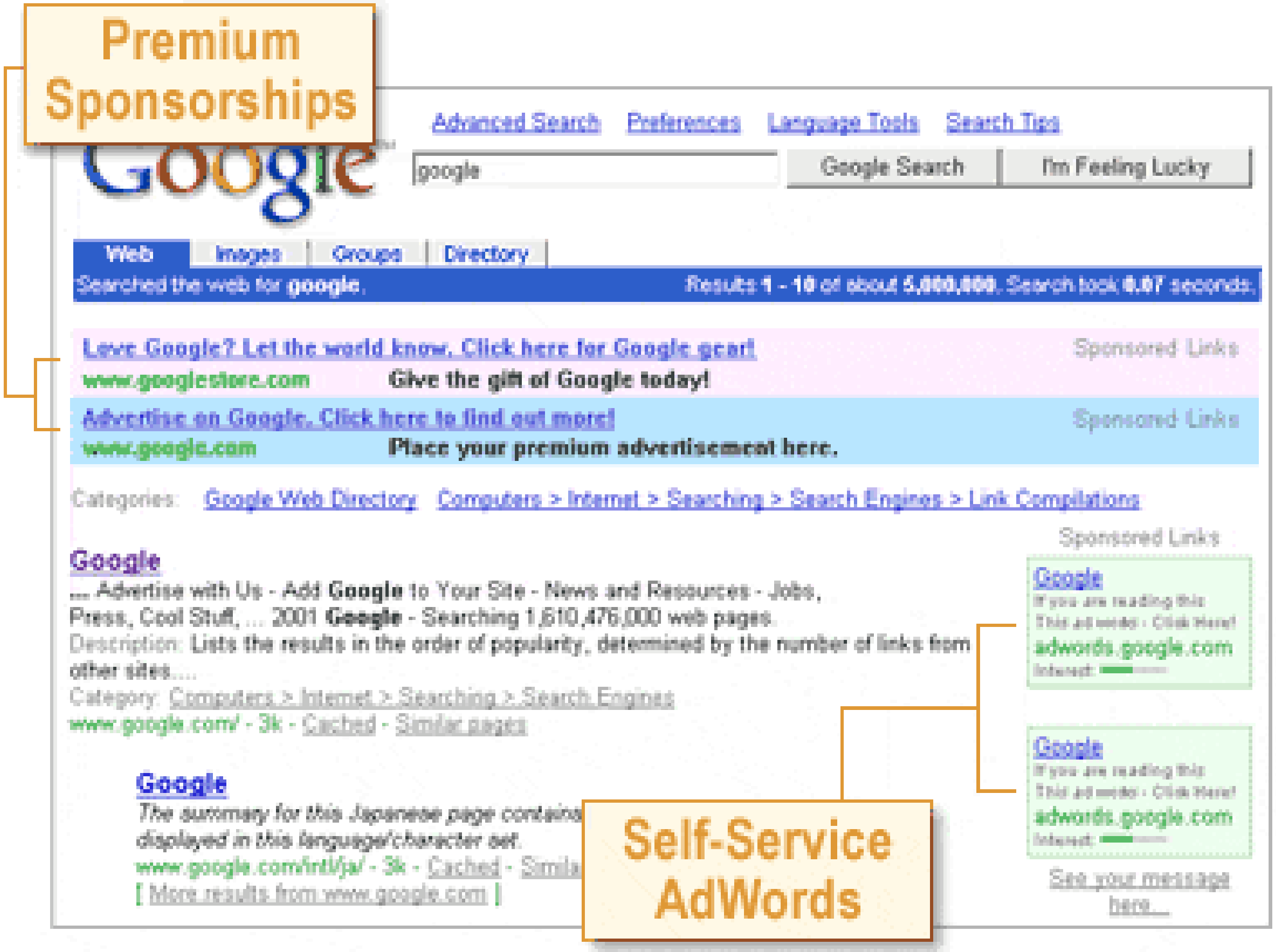
Scalable Search Services:

- 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



Premium Sponsorships

Searched the web for **google**. Results **1** - **10** of about **5,000,000**. Search took **0.07** seconds.

Love Google? Let the world know. Click here for Google gear! Sponsored Links
www.googlestore.com **Give the gift of Google today!**

Advertise on Google. Click here to find out more! Sponsored Links
www.google.com **Place your premium advertisement here.**

Categories: Google Web Directory Computers > Internet > Searching > Search Engines > Link Compilations

Google
... Advertise with Us - Add **Google** to Your Site - News and Resources - Jobs, Press, Cool Stuff, ... 2001 **Google** - Searching 1,610,476,000 web pages.
Description: Lists the results in the order of popularity, determined by the number of links from other sites....
Category: Computers > Internet > Searching > Search Engines
www.google.com/ - 3k - Cached - Similar pages

Google
*The summary for this Japanese page contains ... displayed in this language/character set.*
www.google.com/intl/ja/ - 3k - Cached - Simila...
[ More results from www.google.com ]

Sponsored Links

Google
If you are reading this
This ad works - Click Here!
adwords.google.com
Interest:

Google
If you are reading this
This ad works - Click Here!
adwords.google.com
Interest:

See your message here...

Self-Service AdWords

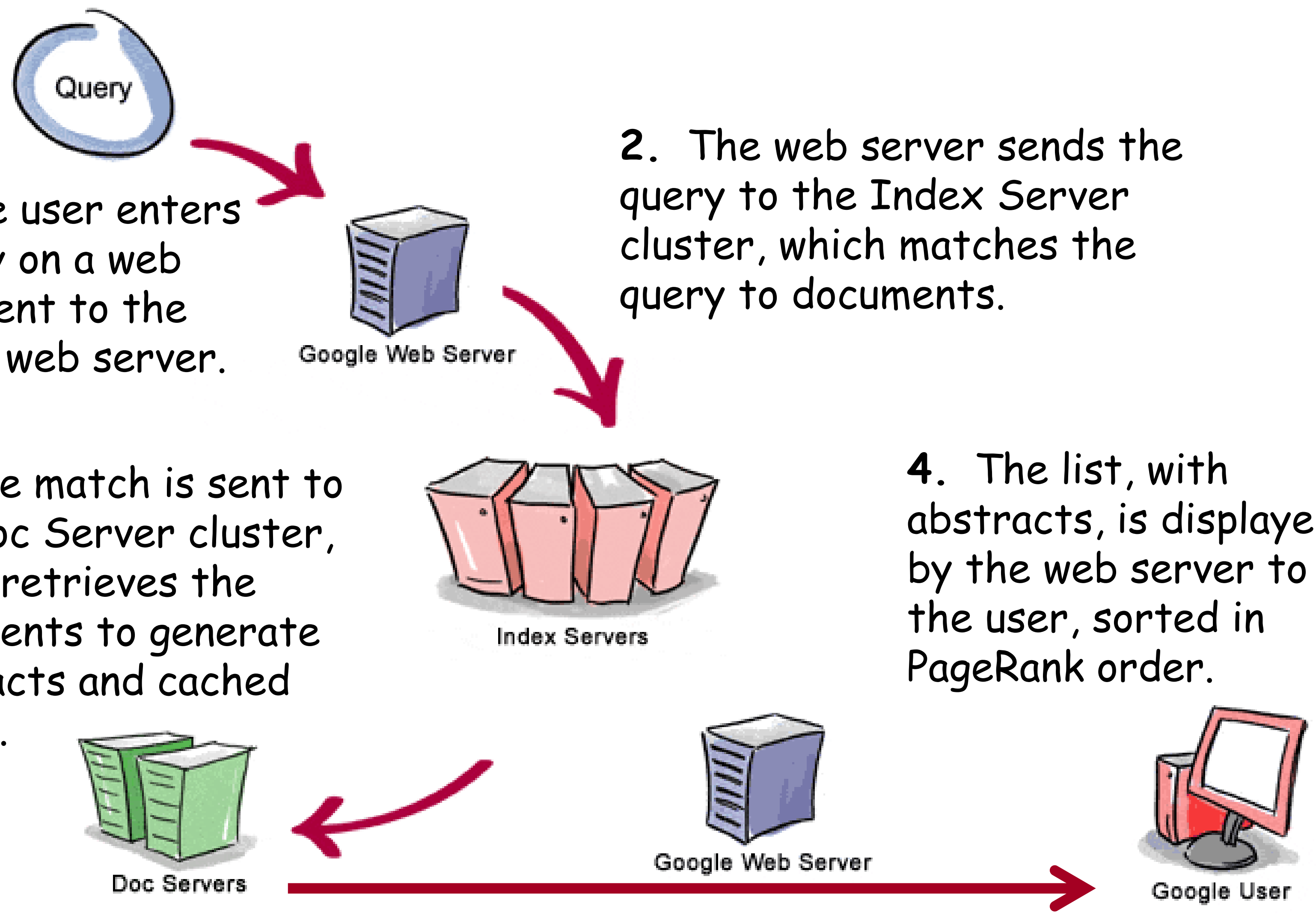
# Technical Highlights

- **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.
  - PageRank results are not modified by sponsors or employees.
- **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!*
- **Bayesian Spelling-Suggestion Program:** Offers suggestions for misspelled words in queries, making searching easier.
  (*"Did you mean…?"*)

# Life of a Query



1. The user enters a query on a web form sent to the Google web server.

2. The web server sends the query to the Index Server cluster, which matches the query to documents.

3. The match is sent to the Doc Server cluster, which retrieves the documents to generate abstracts and cached copies.

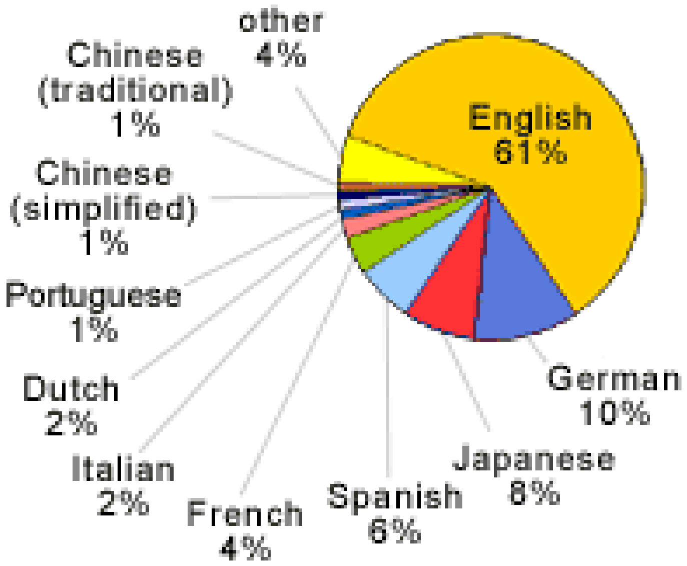
4. The list, with abstracts, is displayed by the web server to the user, sorted in PageRank order.

# Searching Habits

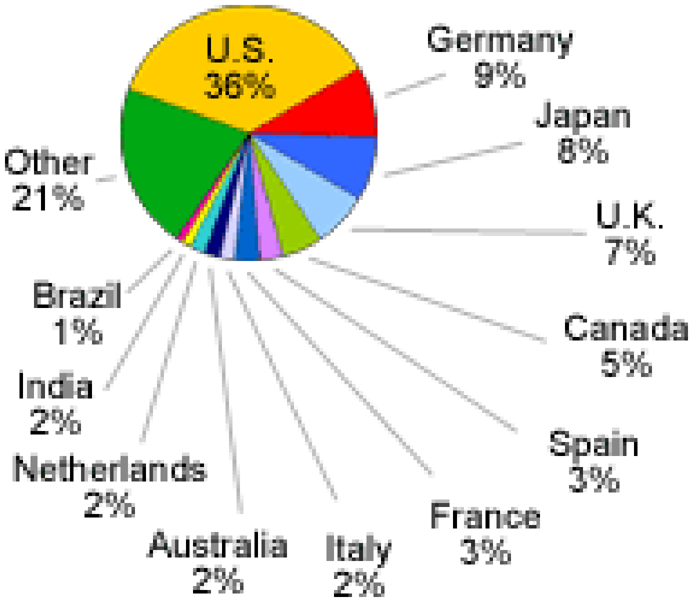
Google's *Zeitgeist* has interesting statistics about people's searches by logging the search queries!

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

**Languages used to search Google** (October 2001)



**Origin of Google searches by country** (October 2001)



# Searching Habits *(continued)*

## Top Ten Gaining Queries (October 2001)

1. Anthrax
2. Windows xp
3. Al jazeera
4. Milzbrand (anthrax in German)
5. Cipro
6. AC-130
7. Smallpox
8. Harry potter
9. Xbox
10. Michael Jordan

## Top Ten Declining Queries (October 2001)

1. Nostradamus
2. World Trade Center
3. American Flag
4. Nimda
5. Pentagon
6. Cantor Fitzgerald
7. Fantasy Football
8. American Red Cross
9. FBI
10. FAA

## Top Five Gaining Image Queries: (October 2001)

1. Pumpkin
2. Osama Bin Laden
3. Heather Graham
4. Aishwarya Rai
5. Drew Barrymore

# Reading Assignment for November 29, 2001

- Google Press Guide (http://www.google.com/press/guide)
- Google Corporate Overview (including pages linked from here, *e.g.*, "Company Milestones") (http://www.google.com/press/overview.html)
- "How Internet Search Engines Work," HowStuffWorks.com (http://www.howstuffworks.com/search-engine.htm/printable)