Search Engines:
Technology, Society, and Business

Course Summary

Marti Hearst

December 10, 2007
Course Goals

• Gain an interdisciplinary understanding of search engines and related technologies.
  - How they work
  - How they affect communication
  - How they affect business
  - How they are changing our understanding of information and knowledge.

• Make the techy parts understandable for everyone.
Intro to the Internet & WWW

- Prof. Hearst

Internet Addresses

- The internet is a network on which each computer must have a unique address.
- The Internet uses IP addresses; for example, herald’s IP address is 128.32.226.90.
- Internet Protocol version 4 (IPv4) supports 32-bit dotted quad IP address format:
  * Four sets of numbers, each set ranging from 0 to 255.
  * UC Berkeley’s LAN addresses range from 128.32.0.0 to 128.32.255.255.
  * Other addresses in the 5865 LAN include 128.32.226.49.
- Using this setup, there are approximately 4 billion possible unique IP addresses.
- Router software knows how to use the IP addresses to find the target computer.
Dr. Jan Pedersen

- The Four Dimensions of Search Quality

- Problem:
  - Ensure that what is indexed correctly reflects current state of the web
  - Impossible to achieve exactly
    - Revisit vs Discovery
  - Divide and Conquer
    - A few pages change continually
    - Most pages are relatively static
Dr. Daniel Russell

- User Experience Issues in Web Search
Class Exercise

- Students as web pages and a search engine
  - Web pages:
    - Web site = where you live
    - Hyperlinks = who you know in class
    - Web page = Beatle’s song title

Unit 2

Jane Tran

I Wanna Hold Your Hand
Dr. Hal Varian

- Search advertising

Factors affecting revenue

\[
\text{Monetization (RPM)} = \frac{\text{Revenue}}{\text{Queries}} \times \left( \frac{\text{1K}}{} \right) = \frac{\text{Revenue}}{\text{Clicks}} \times \frac{\text{Clicks}}{\text{Queries}} = \frac{\text{Revenue}}{\text{Clicks}} \times \frac{\text{Queries w/ Ads}}{\text{Queries}} \times \frac{\text{Ads}}{\text{Queries w/ Ads}} \times \frac{\text{Clicks}}{\text{Ads}} = \text{CPC} \times \frac{\text{Coverage}}{\text{Depth}} \times \frac{\text{CTR per Ad}}{\text{Price}} \times \text{Quantity} \times \text{Quality}
\]
Dr. Mark Najork

- Web Spam

Examples of synthetic content

- Monetization
- Random words
- Well-formed sentences stitched together
- Links to keep crawlers going
Chris Hoofnagle

- Privacy and Online Information

Search Strings

- AOL releases 20m queries based on 600k users to help researchers
  - Were trying to make routine access more efficient
- Users are uniquely enumerated
- Some easy to identify
  - Users vanity searched name, SSN
- Many others identifiable based on searches unrelated to PII
  - Thelma Arnold, Lilburn, Ga
Dr. Lynn Wilcox

- Multimedia Search
Jason Schultz

- Search and Intellectual Property
Dr. Jaime Teevan

- Personalization and Search

Ranking Results for Re-Finding
Natural Language Processing

- Prof. Hearst

How can a machine understand these differences?

- Get the cat with the gloves.
John Battelle

- The Search

Search and Culture

- The Realization: My God....Google Knows What We Want...
- The Database of Intentions
- Ephemeral to Eternal
- First Use Case: Paid Search
What was most surprising?
Final Projects

• Turn them in using online link
  • http://courses.ischool.berkeley.edu/i141/f07/assignments/projects.html

• HARD DEADLINE!

• Due Sunday Dec 16, 9am

• Feel free to turn yours in early!
Course Evaluations

• This is the iSchool form
  ■ First page is instructor evaluation
  ■ Back of page is course evaluation

• I’ve also added another sheet with a few specific questions about future course format.

• Instructor does NOT see these until after she turns in the grades.

• Turn in the form to a TA, who will then check you off for attendance.
  ■ TAs will not accept forms until 10 minutes after they are distributed.
Let’s Thank Our TAs!

Eun Kyoung Choe and Ani Sen
Thank you!

And Happy Searching!