The Web: Concepts and Technology

January 15: Course Overview
Today’s Plan

- Who am I?
- What is this course about?
- Logistics
- Who are you?
Meet Your Instructor

- **Instructor:** Eugene Agichtein  
  Web: [http://www.mathcs.emory.edu/~eugene](http://www.mathcs.emory.edu/~eugene)  
  Email: eugene@mathcs.emory.edu  
- **Office Hours:** TBD  
- **Office:** Emerson Hall E500  
- **Telephone:** (404) 727-7962

**TA:** TBD
E500 Emerson
About the instructor: Prof. Eugene Agichtein

- Sept 2006-: Associate Professor in Math & CS
  - Undergraduate courses taught: CS170 x 2, CS171 x 4, CS190 x 2, CS370 x 1

- Summers 2007-2012: Consulting at Yahoo! Research, eBay, Microsoft Research, Google

- 2004 to 2006: Postdoc researcher at Microsoft
  Web search, text & data mining

- 1998-2004: Ph.D. in Computer Science from Columbia University: Scalable information extraction

Research: Developing Intelligent Systems to Help People Find Information Online

Search, browsing behavior

User-generated content, social networks

Also interested in: human-computer interaction, cognitive information seeking models, visual attention.
Meet everyone in the class

• I, <state your name>

• Major(s), Year at Emory, programming experience

• If you had $1,000,000 to invest and unlimited time, what software/app would you want to build?
Course Topics (rough chronology)

- **Web browsing** and other usability issues
- **Web infrastructure**: history, networking/protocols, hardware, energy
- **Web services**: mash-ups (usage) and underlying technology
- **Web Search**: history, algorithms, structure of the web, technology (implementation)
- **Web applications**: E-commerce, advertising
- Abuse: spam, hacking and the gray areas
- **Web 2.0** and online social networks theory: influence, search, technology (implementation)
- **Web 2.0** artifacts: (Blogs, Wikipedia, Yahoo! Answers)
- **Web 2.0** development: Facebook development platform
What is the Internet?

- The largest network of networks in the world.
- Uses TCP/IP protocols and packet switching.
- Runs on any communications substrate.

From Dr. Vinton Cerf, Co-Creator of TCP/IP
Structure of the Internet
Brief History of the Internet

- 1968 - DARPA (Defense Advanced Research Projects Agency) contracts with BBN (Bolt, Beranek & Newman) to create ARPAnet

- 1970 - First five nodes:
  - UCLA
  - Stanford
  - UC Santa Barbara
  - U of Utah, and
  - BBN

- 1974 - TCP specification by Vint Cerf

- 1984 – On January 1, the Internet with its 1000 hosts converts en masse to using TCP/IP for its messaging
Web Search: Google

1997

2000
Google Architecture

- **URL Server**
  - sends lists of URLs to crawlers
- **Crawler**
  - downloads web pages
- **Store Server**
  - compresses & stores web pages into the repository
- **Indexer**
  - reads the repository & uncompresses the documents
  - parses the documents
  - creates forward index
  - parses out the links
- **URL Resolver**
  - converts relative URLs to absolute URLs and then to docIDs
  - generates a database of links
  - puts the anchor text into the barrels
- **Sorter**
  - generates the inverted index
- **Searcher**
  - answers queries
Web Search: Google (continued)

2001

2007
Search ads

Promoted ad

Search results

RHS ads
How does Google work?

• Advertiser bids on search terms
  – Can use “Traffic Estimator” to see how many clicks they will get
  – Can choose “exact” or “broad” match

• Ordering
  – Overture: high bidder gets first position, $2^{nd}$ highest bidder second position, and so on, with exact match first.
  – Google: rank by bid $\times$ predicted CTR. Puts best ads on top of page.
Biggest auction in the world

- According to Comscore, there were 4.8 billion web searches worldwide July 2005
- Roughly half of those searches displayed ads, so that’s 2.4 billion auctions per month
- This is a very conservative estimate...
Ads system from viewpoint of seller

- Choose your creative text
  - Standardized size
  - Various tips available to aid choice
- Choose your keywords
  - Exact, broad, phrase, negative keyword
  - Follow ad guidelines
- Implement conversion tracking if desired
Users learn to ignore ads!

- Heat map:
  Detect gaze position and duration using eye tracking

- “Box Blindness”
AdSense Program

• “Content ads”
  – Use keyword matching engine to pick ads related to content on page
  – Matching algorithm tweaked to deal with multiple content on page
  – Auction needs tweaking to deal with position effects
  – Generally lower CTRs, lower conversion performance, so adjustments made in payment to compensate for performance difference
Technology

- Engine to match ads to queries
  - Choose most relevant ads using various criteria (plurals, synonyms, elimination of duplicates, etc.)
  - Returns up to 10 ads to show on a page
  - Promotion policy: depends on CTR performance and CPC. Best ads get best position.

- Auction engine
  - Ranks ads by CPM, records clicks, payments, etc.
This was “surface” web
The Invisible, Deep, or Hidden Web

- Web sites or information that Google or other popular search engines are not fully indexing
- Websites specifically excluded by the search engine
Spam 101

kaiser pharmacy online
Spam 101

Save today on Viagra, Lipitor, Zoloft, ...

Phentermine 90 Pills/$119
Spam 101

**Spamming** = misleading search engines to obtain higher-than-deserved ranking

Techniques

- Spamming
- Hiding
  - Term
  - Link
  - Content Hiding
  - Cloaking
  - Redirection

**Link spamming** = building link structures that boost PageRank score
Techniques / Hiding

• Content hiding

```html
<style type = "text/css">
  body {
    background-color: white;
    color: white;
  }
</style>

<a href = "..."><img src = "1x1.gif"></a>

<div style = "visibility: hidden">You can’t see me!</div>

• Cloaking
  ▪ Identify web crawlers
  ▪ Serve a different version of the page
Web 2.0: It’s Hard to Define, But I Know it When I See it…

Emerging Tech
- Web Services / API’s
- Folksonomies” / Content tagging
- AJAX”
- RSS

Some Apps You may Know…
- Flickr
- Google Maps
- Blogging & Content Syndication
- Craigslist
- Facebook, Linkedin, Tribes, Ryze, Friendster
- Del.icio.us
- Upcoming.org
- 43Things.com

Major Retailers
- Amazon API’s
- Google Adsense API
- Yahoo API
- Ebay API

"[This is] not my mom's Internet…It's changing, and it's changing because we're looking at the share-shifting—the time people are looking at TV, reading a magazine, listening to the radio—they're not replacing each other; they're coming together." - AOL Exec / May 2005
Web 2.0: Evolution Towards a Read/Write Platform

<table>
<thead>
<tr>
<th>Web 1.0</th>
<th>Web 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretty much HTML pages viewed through a browser</td>
<td>Web pages, plus a lot of other “content” shared over the web, with more interactivity; more like an application than a “page”</td>
</tr>
</tbody>
</table>

**“Read”**

**“Write” & Contribute**

**“Page”**

“Post / record”

**“static”**

“dynamic”

Web browser

Browsers, RSS Readers, anything

“Client Server”

“Web Services”

Web Coders

Everyone

“geeks”

“mass amaturization”

**Primary Unit of content**

**Content Created by...**

**Domain of...**

**Mode**

**State**

Viewed through...
Recommendations

Search

Recommendations

Items

Products, web sites, blogs, news items, …
Well-known recommender systems: Amazon and Netflix
Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!
Course Logistics

- Website: [http://www.mathcs.emory.edu/~eugene/cs190/](http://www.mathcs.emory.edu/~eugene/cs190/)

- Facebook group, Twitter account: TBA
Course Logistics (cont’d)

- Text book: strongly recommended as reference

- All the assigned readings will be available online
  - Online tutorials (Unix, HTML, Javascript)
  - Popular press and scientific articles
  - Occasionally parts of textbook chapters as needed.

- The expected reading amount will be roughly 1 article or (part of) a book chapter per week.
Grading

25%  Individual assignments  5 total, mostly of them early in the semester

30%  Group small projects  2 total, roughly 2 weeks for each one

30%  Group final project  1 total, roughly 3-4 weeks long

10%  Quizzes  6-8 total, roughly every two weeks

5%  Class participation  ask questions, participate in discussions… the usual.
Course Expectations

- Web programming (and any other programming) is a language and a way of thinking. Be ready and willing to learn new things.

- Be prepared to think hard

- A good understanding of high-school level mathematics and logic is helpful.

- Most importantly, I expect that you are interested in learning about the web and the underlying web technologies.
The main point of this lecture

This course will be fun and you will learn alot, but expect to spend the time and effort:

- To learning a new language and way of thinking;
- To spend some sweat developing, experimenting, and tweaking your projects

The reward is a set of useful skills for other courses, life, and employment