Overview of Computer Science

CSC 101 — Summer 2011
Internet / World Wide Web
HTML: The Language of the Web
Web Searching / Google
Lecture 3 — July 8, 2011

Announcements

• Writing Assignment #1
  – Due Monday 7/11
  – Hand in printed out copy to me on Monday
  – Bring electronic copy (ie. MS Word) to lab#1 on Tuesday

Objectives

• Overview and history of the Internet and the World Wide Web
• What’s the difference between the Internet and the World Wide Web?
• How do clients and servers work together to make the Web happen?
• Brief introduction to HTML
• Searching the Internet / Google
The Internet

- The Internet
  - A loose collection of interconnected networks
  - No single entity controls the Internet
  - The Internet is an internet – an interconnected network
    - The “Internet” is a proper noun, and so is capitalized
    - Many kinds of information fly through it constantly
  - No one validates or directly polices this information
    - It is up to you, the user, to determine the trustworthiness and legality of information you get through a network

The Web

- The World Wide Web (WWW)
  - Are the Web and the Internet the same thing?
    - No!
      - The Web is just one of the many kinds of traffic on the Internet
        - The Web is made up of only that information on the Internet that uses a particular method (HTML & HTTP)
      - Your computer is part of the Internet whenever you are connected
      - You will make pages that will become part of the Web during the first few labs

Web Browsers

- Web browsers are programs that make it easy for anyone on the Internet to access information on the Web
- Many to choose from; they all do the same job
  - Firefox (Mozilla)
  - Internet Explorer (Microsoft)
  - Safari (Apple – Mac, iPhone, PC)
  - Opera
  - Chrome (Google – Android)
  - MSN TV [formerly WebTV] (Microsoft)
  - Other devices with browsers

http://marketshare.hitslink.com/browser-market-share.aspx?qprid=0
History of The Internet

• Q: Who invented the Internet?
• Originally a US military project from the late 60’s: ARPANET
  - Many contributors – no single inventor
  - Designed to survive a nuclear attack
• Initially four sites in CA and UT (1969)
  - Expanded into academics and research in the 70’s
  - Separated from MILNET in early 80’s
• ARPANET project decommissioned in 1990
  - Many interconnected private and public networks now make up the Internet
• Originally only non-commercial uses allowed
  - Peer-pressure only, since there was — and is — no central control!

History of The Internet

• The ARPANET in 1971 – 18 sites

History of The Internet

• The ARPANET in 1980 – about 75 sites
History of The Internet

- The Internet today – over 1,000,000,000 nodes worldwide
  - US-based backbones (North America):

  ![Map of US-based backbones](http://www.caida.org/tools/visualization/mapnet/Backbones/)


History of The Internet

- US-based backbones (worldwide)

  ![Map of US-based backbones](http://www.caida.org/tools/visualization/mapnet/Backbones/)


Top-Level Domains

- Six original “top-level domains” (TLDs)
  - .edu – educational institutions
  - .gov – government agencies
  - .org – non-profit organizations
  - .net – organizations that make the network work
  - .mil – military sites
  - .com – commercial companies
  - .int – “international” – the rest of the world – was added later

- TLDs are assigned by ICANN, a US-run non-profit corporation
Top-Level Domains

- **International** top-level domains came later:
  - .us – The United States, .fr – France
  - .ca – Canada, .de – Germany
  - .uk – United Kingdom, .cn – China
  - .jp – Japan, .tv – Tuvalu Islands
  - ... etc.

- More recent **additions**:
  - .aero, .biz, .coop, .info, .jobs, .mobi, .museum, .name, etc.
  - **Internationalized TLDs** (non-Roman characters) now available, such as:
    - .中国 (China); .السعودية (Saudi Arabia); .俄 (Russia); .नेपाल (Nepali)

History of The World Wide Web

- **Invented by Tim Berners-Lee at CERN** (1989)
- Originally designed for distributing scientific research results
  - Text pages that could be shared among different computer systems
  - Simple, text-based browsers
- Quickly adopted by other organizations

History of The World Wide Web

- First graphical browser (Mosaic) developed in 1993 by an undergrad at University of Illinois (Marc Andreesen)
  - Distributed freely
  - Widely used in academics and government
  - Andreesen co-founded Netscape to commercialize Mosaic
    - Internally they called it “Mozilla”
    - Mozilla 1.0 = a big, gory version of “Mosaic”?!
    - Originally still distributed without charge
  - Microsoft’s Internet Explorer came later
  - Netscape now commercialized (AOL)
  - Mozilla / Firefox still free (split off from Netscape)
  - Other competing browsers also available
  - No matter which browser is used, it’s still the same Web
History of The World Wide Web

- “Ban” on commercial traffic on the Web lifted soon after Mosaic released
- .com quickly becomes largest domain
- E-commerce explosion starts in the late 1990’s
  - Amazon.com founded 1994 by Jeff Bezos
  - IPO 1997 @ $18/share
  - Each share bought then now worth ~$400
  - 2200% return over 9 yrs
  - First profit in 2001
  - Now over $10,000,000,000/yr in sales
  - Jeff Bezos’ net worth ~$4,300,000,000

Growth of the Internet and the Web

<table>
<thead>
<tr>
<th>Year</th>
<th>Internet Nodes*</th>
<th>Web Servers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>5,089</td>
<td>–</td>
</tr>
<tr>
<td>1984</td>
<td>1,024</td>
<td>–</td>
</tr>
<tr>
<td>1986</td>
<td>992,000</td>
<td>–</td>
</tr>
<tr>
<td>1988</td>
<td>56,000</td>
<td>–</td>
</tr>
<tr>
<td>1990</td>
<td>235,000</td>
<td>–</td>
</tr>
<tr>
<td>1992</td>
<td>1,024,000</td>
<td>–</td>
</tr>
<tr>
<td>1994</td>
<td>12,881,000</td>
<td>300,000</td>
</tr>
<tr>
<td>1996</td>
<td>16,195,000</td>
<td>18,170,000</td>
</tr>
<tr>
<td>1998</td>
<td>36,739,000</td>
<td>4,279,000</td>
</tr>
<tr>
<td>2000</td>
<td>93,969,000</td>
<td>16,170,000</td>
</tr>
<tr>
<td>2002</td>
<td>162,129,000</td>
<td>127,962,000</td>
</tr>
<tr>
<td>2004</td>
<td>239,611,000</td>
<td>186,728,000</td>
</tr>
<tr>
<td>2006</td>
<td>439,286,000</td>
<td>106,875,000</td>
</tr>
<tr>
<td>2007</td>
<td>489,774,000</td>
<td>121,061,000</td>
</tr>
<tr>
<td>2008</td>
<td>570,938,000</td>
<td>186,728,000</td>
</tr>
<tr>
<td>2009</td>
<td>681,065,000</td>
<td>239,611,000</td>
</tr>
</tbody>
</table>

* A “node” is just a computer connected to the network.
(They can only be counted if they respond to a “ping”, so these numbers are too low.)

Sources: (1) www.isc.org/ds; (2) http://news.netcraft.com/archives/web_server_survey.html
The Client / Server Paradigm

- An efficient way to communicate information is to use clients and servers
- Clients are nodes that:
  - Request information
  - Receive information
  - Display information
- Servers are nodes that:
  - Store information
  - Process requests from clients
  - Send information to clients in response to requests
- Peers are nodes that can act as either or both
- A common protocol and language are required so that clients and servers can communicate

Web Browsers and Web Servers

- Web browsers are clients
  - Request, receive, and display Web documents
- Web sites are servers
  - Store Web information
  - Process browser requests and send information back to the browser
- The protocol used is HTTP
  - HyperText Transfer Protocol
    - Method by which browsers and servers communicate
- The language used is HTML
  - HyperText Markup Language
    - The actual description of the content and layout of the page
- The addresses used by clients to find servers are URLs
  - Uniform Resource Locator
URLs

- Browsers request a page by its URL
  - Uniform Resource Locator
- Example:
  
  http://www.wfu.edu/pryoree/index.htm

- Other protocols also supported by most browsers:
  
  https: ftp: file: etc.

Web Publishing

- HTTP defines how browsers and servers communicate
  - A protocol is a set of rules for having a conversation
- HTML is the language they use
  - A conversation needs a language that is understood by both sides of the conversation
  - HTML is a language that is used to define both the content (text, images and other elements) and the structure (layout and appearance) of Web pages

Web Publishing

- HTML — Hypertext Markup Language
  - “Hypertext”
    - Text documents containing hyperlinks (“links”) to other documents
    - Cross-references to a variety of information in various locations, local or remote
  - “Markup”
    - Originally from print publishing:
      - Handwritten instructions to the typesetter
        - Text size, bold or italic, headings vs. body text, column layout, etc.
  - “Language”
    - Standardized set of instructions for a computer
Web Publishing

- HTML was designed to be:
  - Platform independent
    - Any computer, any browser – the page should look about the same
  - Resolution independent
    - Any screen or printer size or quality – the page should look the same
  - Multimedia
    - Text, images, video, sound, interactive tools, etc., can be combined on each page
  - Hyperlinked
    - Immediate access to other content, whether local or remote
  - Flexible
    - Allows for the addition of new, unexpected technologies, enabling the great growth of the Web

Creating Web Pages

- HTML files are plain text with markup tags
  - Can be created and edited as plain text
    - Example: Notepad
  - HTML editing programs can make this easier
    - Example: Dreamweaver
    - WYSIWYG editing, like a word processor
  - Site Managers help track the files that make up a site
    - Example: Dreamweaver’s “Site” menu
    - Helps organize and track all the files
    - HTML files, images, programs, etc.
    - Helps keep the Web server files in sync with local files

HTML

- HTML consists of plain text with markup tags
  - Plain text: standard text characters
  - Markup tags: embedded commands or codes (within the plain text) that describe
    - Content (what is on the page)
    - Structure (layout or appearance of the page)
- Example:
  
  ```html
  <p>This is some <strong>sample</strong> HTML.</p>
  ```
HTML Tags

- Tags are not case-sensitive
- Most tags are “paired tags”
- Example:

  ```html
  <p>This is some <strong>sample</strong> HTML.</p>
  ```

  Appears as:

  This is some sample HTML.

Searching the Web

- Specialized sites for searching the Web:
  - *Internet search engines*
    - Search engine web sites help to index the Web
    - But, the Web is so huge that it’s impossible to index everything
    - Both manual and automatic methods can be used to find and index pages

Search Engines

- Manual selection of web sites
  - Used to create a “directory” of Web sites
    - Yahoo! started out this way
  - Pages examined by people who classify and categorize them
    - Organized into topics, geographic areas, etc.
    - Human evaluation is made before a site is included
  - No longer the main method for indexing Web sites
    - The Web is too huge, so automatic methods are necessary
Search Engines

• Automatic indexing of the Web
  – Web sites are found by special software called a robot, web crawler or spider
  – Runs forever on the Web, looking for links
  – Follows all links that it can find in every page it finds
  – Every page it finds is added to its list
    – Some filtering and automatic evaluation
    – Relevance ranking based on content
    – Keywords from the page are indexed
  – Forms very large and comprehensive lists with some relevance ranking
  – Good search engines do this very well

Search Engines

• Most search engines now use automatic techniques with constant improvement
  – Humans continuously monitor the quality of results and improve the automatic methods
  – A great history of search engines can be found at SearchEngineHistory.com
  – Many old versions of Web pages can be found at the WayBack Machine: www.archive.org

Search Engine Popularity

• Many engines to choose from
• Popularity depends on
  – How well they’re marketed
  – How well they work

<table>
<thead>
<tr>
<th>Engine</th>
<th>1/02</th>
<th>1/04</th>
<th>7/05</th>
<th>7/07</th>
<th>4/08</th>
<th>12/08</th>
<th>5/09</th>
<th>7/10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>24%</td>
<td>25%</td>
<td>36%</td>
<td>46.2%</td>
<td>53.3%</td>
<td>62.0%</td>
<td>64.2%</td>
<td>61.6%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>32%</td>
<td>22.5%</td>
<td>23.8%</td>
<td>20.1%</td>
<td>17.5%</td>
<td>16.8%</td>
<td>15.8%</td>
<td>20.1%</td>
</tr>
<tr>
<td>MSN</td>
<td>14%</td>
<td>12.6%</td>
<td>9.6%</td>
<td>13.6%</td>
<td>9.7%</td>
<td>9.8%</td>
<td>10.3%</td>
<td>12.6%</td>
</tr>
<tr>
<td>AOL</td>
<td>n/a</td>
<td>5%</td>
<td>5.4%</td>
<td>6.3%</td>
<td>5.2%</td>
<td>4.3%</td>
<td>4.1%</td>
<td>3.7%</td>
</tr>
<tr>
<td>Ask</td>
<td>n/a</td>
<td>5.4%</td>
<td>1.6%</td>
<td>2.6%</td>
<td>2.7%</td>
<td>2.1%</td>
<td>2.0%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Other</td>
<td>n/a</td>
<td>15%</td>
<td>11.7%</td>
<td>8.5%</td>
<td>6.0%</td>
<td>4.4%</td>
<td>4.4%</td>
<td>3.9%</td>
</tr>
</tbody>
</table>

US Data. Sources: comScore via SearchEngineWatch.com
Google

- Google is the search engine with the largest market share by far
  - What makes them so great?
    - Comprehensive, clever indexing of the web
    - Lots of advanced technologies
    - Available in 130 languages and localized to 175 countries
    - More and more tools and features available
    - The Xerox effect

1995: Two CS graduate students at Stanford got together
- Sergey Brin and Larry Page
- 1998: Founded Google.com in a friend's garage
- Maxed-out their credit cards to fund it
- Quickly started making lots of money
- 2004: Went public
  - Raised $20,000,000,000
  - Now worth $150,000,000,000 ... with $5,000,000,000 cash in the bank
- Sergey and Larry now tied for #16 on the top 400 list
  - Worth $13,000,000,000 each
  - They each sell off a few hundred million in stock each month

What makes Google such a good search engine?
- Search results are based on PageRank
  - A web page's relevance is judged based on "voting" -- how many other pages point to it
- Very good text matching
  - Gives greater importance to words found in close proximity
  - Many clever text matching algorithms used
- Very fast computers
  - Extremely large compute clusters
    - Over half-a-million dual-processor computers...
• Google still growing
  – Huge data center opening in Lenoir, NC
  – Just one of many worldwide

• Google is expanding into many other areas
  – Android operating system for handhelds
  – Chrome OS operating system for PCs
  – Browsers (Chrome)
  – Desktop tools
  – Mapping (e.g. Google Earth, Google Mars)
  – Calendar, word processor, spreadsheet
  – Email, instant messaging, VOIP
  – WiFi networking
  – Images and video (can search on closed captions)
  – Library (print.google.com)
  – Translations
  – Etc., etc., etc.

• But, they are relatively secretive of their long-term plans
  – Google owns or has obtained rights to more network bandwidth than any other organization, and they are building gigantic data centers all over the country (including a number in the Carolinas)
  – What are they planning to do with this capacity?
  – Google appears to be positioning itself to try to transform all computing, communication, and entertainment from local applications and traditional media to Internet applications (“Cloud computing”)

Google’s mission is to
“organize the world’s information and make it universally accessible and useful”

Google’s motto is:
“Don’t Be Evil”
- Are there risks involved?
  - Immediate access to lots of data
  - Without oversight or guidance about the value of the data
  - Concerns about privacy
    - Private information collected
      - Which sites your computer has visited
      - Search terms used
      - Search results can depend on
      - Geographic location
      - Browsing history
    - Ads on Gmail and other Google tools are keyed to content
    - Faces, license plate numbers, etc. visible on Street View
    - Storage of emails and documents on Google’s central servers