

Lecture

- Questions about Project #3?

Web Site Validation

- Finding who owes a machine on the internet is important.
 - dhmo.org (who is responsible for it?)
- There are several databases (whois servers) which keep track of owners of various domains on the internet
- Unix whois command provides information about domains
 - Domain name
 - Registrar – (organization that registered the domain)
 - Whois server – where you can find information about the domain
 - When using whois provide domain names (not machine names)
- Example using cs.umd.edu and the Unix command
 - Execute the Unix whois command to find the whois server to contact
 - Execute the Unix whois command with the `-h` option

ICANN

- ICANN - Internet Corporation for Assigned Names and Numbers
- Responsibilities
 - Ensures every IP address is unique
 - Users can find valid addresses
 - Oversees distribution of IP addresses and domain names

Internic

- www.internic.net –
 - Provides information regarding Internet domain name registration services.
 - default whois server that tells you which database contains information about a specific domain.
- Default whois server can only provide information for .aero, .arpa, .biz, .com, .coop, .edu, .info, .int, .museum, .name, .net, or .org
- Looking information for
 - umd.edu, cnn.com
- whois server for .mil domains – whois.nic.mil
- Whois server for .gov domains – whois.nic.gov
- <http://www.internic.net/faqs/domain-names.html> - Provides great information about domain names, ICANN and registrars.

Search Exercises

- ❖ Look for a particular term in only a particular site
- ❖ Look for sites with a particular term in the title
- ❖ Look for all sites with links to a particular site
- ❖ Look for the definition of a term
- ❖ Look for information google keeps about a site
- ❖ Look for a site with all the query terms in the title
- ❖ Look for a site with all the query terms in the url

Software Agents (“spiders”)

- ❖ Search engines need to build a database with a collection of links to web pages.
- ❖ To find pages on the web search engines sites use the so-called software agents.
- ❖ A software agent – program that traverses the web automatically.
- ❖ Software agents are also refer to as “spiders”, “robots”, “web crawlers”, “wanderers”. How crawlers work:
 - ❖ They analyze a page for hyperlinks
 - ❖ They follow the hyperlinks and repeat the process

Software Agents (“spiders”)

- ❖ Spiders use most of the text on a page as keywords
- ❖ It can assign different weights to words found in the title of the page.
- ❖ Unimportant words (e.g., “the”, “or”, etc.) can be ignored.
- ❖ Unimportant words are called “stop words”.
- ❖ Most search engines allow you to submit an URL for its agents to visit .
- ❖ Spiders don’t move from site to site. The search process is performed via requests submitted by the search engine server.
- ❖ A “Search Engine” looks through the documents gathered by a robot.

Example of Robots

- ❖ **Googlebot: Google's Web Crawler –**
 - ❖ www.google.com/bot.html
- ❖ You can request googlebot not to index parts or all your site.
- ❖ A list of web robots can be found at:
 - ❖ www.robotstxt.org/wc/active.html

Resource Type Identifiers

- ❖ Remember that an URL specifies the type (protocol) of resource you are interested
 - ❖ telnet
 - ❖ gopher
 - ❖ ftp or file – request a file via anonymous ftp
 - ❖ wais – connect to a gives WAIS server (Wide Area Information Server)
 - ❖ https – secure connection to secure http server
- ❖ A web browser client will connect to the given host via the appropriate port based on the type of resource.

Other URL Resource Types

❖ Accessing Internet newsgroups

- ❖ news or nntp (Net News Transfer Protocol)

- ❖ Example: `news://nntp.server.name/news.group.name`

- ❖ news could be replaced by nntp (the syntax depends on the browser)

❖ Mail resource

- ❖ `mailto:user@domain.name`

❖ Instant messaging

- ❖ `aim:goim?screenname=buddy&message=content`

Identifying a file's type

- ❖ A file extension is used by a browser to “guess” how the file should be displayed.
- ❖ A browser can be customize to display different types of files

<u>❖ Suffix</u>	<u>File type</u>
❖ .txt	text file
❖ .html	hypertext document
❖ .shtml	html document wit server-side processing being done
• .gif	GIF image
• .jpg	JPEG image
• .ps	PostScript file
• .au	AU format sound file
• .wav	WAVE format sound file
• .ram	RealAudio audio/video format (typically streaming)
• .mp3	MP3 compressed sound file
• .mpg	MPEG movie (more than one mpg level possible)
• .avi	AVI movie (though many codecs for avi format family)
• .zip	ZIP compressed file

Downloading Programs

- ❖ A browser can download a program that can be executed in the client machine
- ❖ plug-ins – products that allow a program to run within a browser
- ❖ Java and Flash programs are very popular