WEB ENGINEERING & DEVELOPMENT

SWE 363

Spring Semester 2008-2009 (082)

Module 1-1-2:
Internet Basics for Web Development

Dr. El-Sayed El-Alfy
Computer Science Department
King Fahd University of Petroleum and Minerals
alfy@kfupm.edu.sa

Objectives/Outline

- Module Objectives
  - Learn the basics of the Internet and the Web
  - Identify and describe the key elements of the Internet and the Web
- Lecture Outline
  - URL Structure
  - Finding Information on the Web
    - Search Engines
    - Other means
  - Web 2.0
  - Questions & Answers

URL Structure

- Uniform Resource Locator (URL)
  - Represents the address of a resource on the Web
- URL defines:
  - Protocol used to access/transfer the document (such as HTTP or FTP; the default is HTTP)
  - Server that hosts the document and its domain name
  - Protocol port number of the server (optional; the default is 80)
  - Path and document name (the default is index.html)
- General form of URL

  protocol: / /server.domain_name:port/item_name
- Example

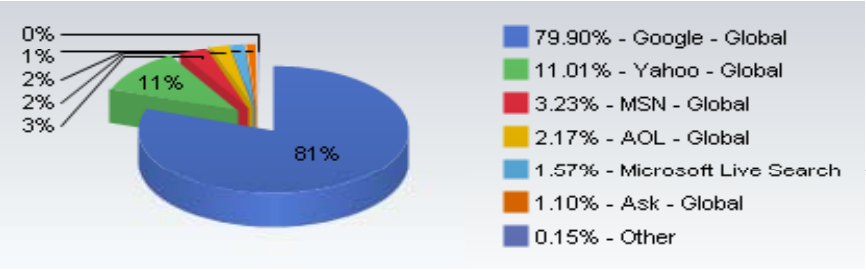
  http://www.kfupm.edu.sa/dad/links.html

Search Engines

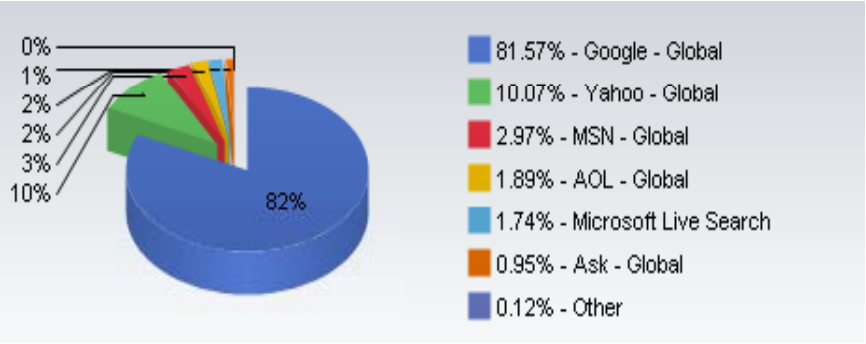
- The Web provides a wealth of information on almost any topic – huge volume of online content
  - No central catalog is possible
- Search engines are the primary tools used to help find relevant information on a specific topic
  - search engine = an information retrieval system that helps find information stored on a computer system, such as the Web, that is relevant to a user query
- There are many search engines
  - Some of them are used for general search but others (called vertical search engines) focus on specific topics (e.g. bioinformatics, medical, Job, Business, real estate, travel, etc)
  - Examples of major search engines
    - Google, Yahoo, MSN, Infoseek, AltaVista, Ask.com, Excite, etc.
    - See a list of them at http://en.wikipedia.org/wiki/List_of_search_engines
  - Differ in their capabilities and the way they work

Search Engines …

- Total market share of major search engines



Oct 2008

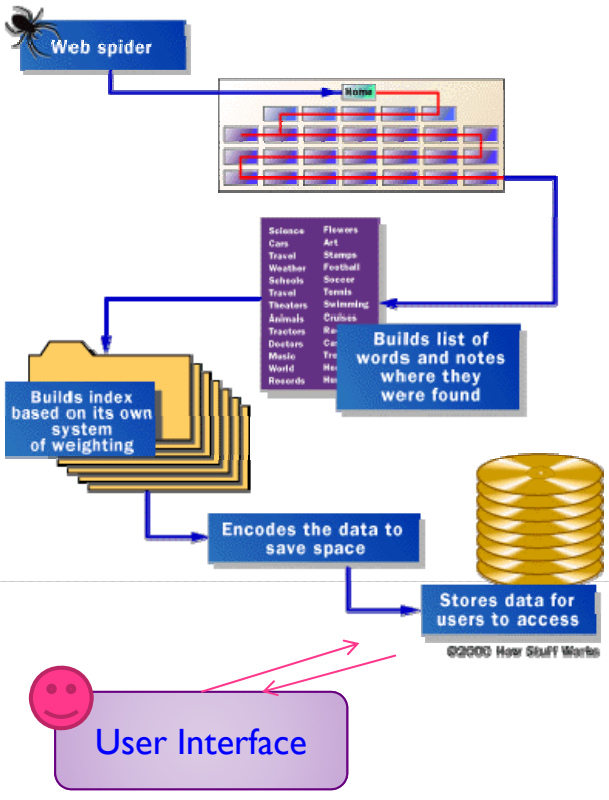


Feb 2009

[Source: Net Applications]

How Web Search Engines Work

- Web crawling (spider)
  - Navigates the web, retrieves web pages that satisfy certain criteria
  - Starts with a popular Web site containing lots of links, such as Yahoo then continues until it finds a logical stop, e.g. a dead end with no external links or reaching a number of levels inside the Web site's structure
- Indexing
  - Pages are analyzed and a list of words and notes (extracted from titles, headings and other special meta tags) are stored in indexes to facilitate quick information retrieval
- User interface
  - Web-Based GUI
  - Allows users to enter their search criteria (search query)
- Searching
  - A search engine stores information about web pages in a database
  - Records that best match the search criteria are returned to the user



How Web Search Engines Work …

- Large search engines, such as Google, index tens to hundreds of millions of web pages involving a comparable number of distinct terms, and answer tens of millions of queries every day.
- Some search engines pre-process the user query to improve the retrieval performance (a process called *query expansion*), e.g.
  - Removing spelling errors
  - Searching for synonyms of the specified keyword
  - Stemming the given keywords to find all morphological forms
- Some search engines *rank* (sort according to relevance) pages that satisfy the criteria specified in the user query

Enhancing Site Visibility

- Search engine optimization (SEO)
  - Developing or tuning a website to improve its ranking in non-paid search engines; to maximize traffic to this site
- SEO can be done using
  - White hat SEO – refers to methods approved by search engines (i.e. do not attempt to deceive search engines), e.g.
    - Offering quality content
    - Using proper metadata and effective keywords
    - Having inbound links from relevant high-quality pages
  - Black hat SEO (spamdexing) – methods that are used to deceive search engines, these can result in temporal improvement, e.g.
    - Googlebomb (or link bomb) is a black hat SEO that attempts to trick the search engine to promote a certain page

Challenges Faced by Search Engines

- Size of the Web
  - Contains more than 3 billion documents, growing very fast and not indexed in any standard vocabulary
- Currency
  - Many Web pages are updated frequently, which forces the search engine to revisit them periodically.
- Relevancy
  - Because the queries one can make are currently limited to searching for key words, may result in many false positives
  - Better results might be achieved by using a proximity-search option or using organic search engines.

Challenges Faced by Search Engines …

- Problem with dynamically-generated Web sites
  - Because these sites may be slow or difficult to index, or may result in excessive results, perhaps generating 500 times more Web pages than average.
- Search engines can be tricked
  - To return pages, in favor of the trick makers, which contain little or no information about the matching phrases.
  - Making the more relevant Web pages pushed further down in the results list
- Indexing secured pages
  - Content hosted on HTTPS URLs pose a challenge for crawlers which either can't browse the content for technical reasons or won't index it for privacy reasons.

[Source:Wikipedia]

Other Ways to Find Info. on the Web

- Meta-search engines
  - have no databases or indexes but they use multiple other search engines and aggregate their results, e.g. WebCrawler
- Web directories
  - human- edited databases that store links in a categorized manner and information about these links (e.g. Yahoo! Directory, Business.com, etc); they can also be automatically created by mining the output of some search engines
- Web portals
  - large multi-service web sites that provides a single point of access to a variety of content and core services.
  - often includes customizable pages, calendars, discussion groups, announcements, reports, searches, email and address books, and access to news, weather, maps, and shopping, as well as bookmarks.
  - often organizes information into channels (customizable page containers) where specific information or an application appears to facilitate locating information of interest by content category.
  - Several universities uses it to create virtual campuses

Web 2.0

- Coined in 2003 by Dale Dougherty at O'Reilly Media to describe the noticeable shift in how people and businesses use the web and develop web-based applications
- In Web 1.0, companies and advertisers produce content for users to access – "brochure web"
- Web 2.0 provides collaborative community-based platforms that involve more user participation, interaction and community contributions
  - Users create content, help organize it, critique it, update it, etc
  - Users create open source software and make it available for anyone to use and modify
  - Users direct how media is delivered and which news and information outlets to trust
  - e.g. Wikis, YouTube, Flickr, MySpace, Facebook, LinkedIn, Google, etc
- Web 1.0 is as a lecture and Web 2.0 is as a conversation
- The growth of Web 2.0 can be attributed for some key factors
  - Improvements in hardware: cheaper and faster
  - Increasing memory capacities and speeds at a rapid rate
  - Broadband Internet access
  - Availability of abundant open source software has resulted in cheaper (and often free) customizable software options
  - Reduced cost of failure to start new Web 2.0 companies

Web 2.0

- User Generated Content
  - Allow users to edit existing content and add new information
  - Collaboration can result in smart ideas
  - But, users also might deliberately submit false or faulty information
    - Web 2.0 companies rely on collaborative filtering to help police their sites
    - Let users promote valuable material and flag offensive or inappropriate material
  - Wikis (What I Know Is) and social networks, e.g. Wikipedia, MySpace, YouTube, Facebook, LinkedIn, Second Life, etc
- Blogs ("Web logs")
  - Websites consisting of entries listed in reverse chronological order; exponential growth; bloggers = blog authors
  - Blogs can also now incorporate media, such as music or videos, e.g. Xanga or LiveJournal
- Social Media
  - Allows users to decide which news articles are most significant, e.g. Digg, Reddit
- Social Bookmarking
  - Allows users to recommend their favorite sites, e.g. del.icio.us, ma.gnolia
- Tagging
  - Labeling already existing web content by subject or keywords that allow anyone to locate information more effectively – pushing the content right to the user's desktop
- RSS feeds
  - RSS = Rich Site Summary
  - Allow users to receive new information as it is updated

Q & A

?

Resources

- *Data Communications and Networking*, 4/e. B.A. Forouzan, McGraw-Hill Higher Education 2007. http://www.mhhe.com/forouzan
- The World Wide Web Consortium (W3C)
- The Anatomy of a Large-Scale Hypertextual Web Search Engine, by Sergey Brin & Lawrence Page at Stanford University
- Dive Into Web 2.0, http://www.deitel.com/freeWeb20ebook/