Chapter 5: Searching for Truth: Locating Information on the WWW

Fluency with Information Technology
Third Edition

by
Lawrence Snyder

Searching in All the Right Places

- The Obvious and Familiar
  - To find tax information, ask the tax office
- Libraries Online
  - Many college and public libraries let you access their online catalogs and other information resources
    - Libraries provide online facilities that are well organized and trustworthy
    - Remember that many pre-1985 documents are not yet available online
- Plus Librarians are real live experts

How Is Information Organized?

- Hierarchical classification (like a family tree)
- Information is grouped into a small number of categories, each of which is easily described (top-level classification)
- Information in each category is divided into subcategories (second-level classifications), and so on
- Eventually the classifications become small enough for you to look through the whole category to find the information you need
  - This is a process of elimination as much as choosing appropriate subcategories

Important Properties of Classifications

- Descriptive terms must cover all the information in the category and be easy for a searcher to apply
- Subcategories do not all have to use the same classifications
- Information in the category defines how best to classify it
- There is no single way to classify information

Design of Hierarchies

- General rules for design and terminology of hierarchies
  - Root is usually at the top (branching metaphor)
    - "Going up in the hierarchy" means the classifications becomes more inclusive or general
    - "Going down in the hierarchy" means the classifications become more specific or detailed
    - The greater-than (>) symbol is a common way to show going down through levels of classification

Levels in a Hierarchy

- A one-level hierarchy has only one level of "branching"—no subdirectories
- To count levels, remember
  - There is always a root
  - There are always "leaves"—the categories themselves
  - The root and leaves do not count as levels
- Groupings may *overlap* (one item can appear in more than one category), or be *partitioned* (every category appears only once)
- Number of levels may differ by category, even in the same hierarchical tree

How Is Web Site Information Organized?

- Homepage is the top-level classification for the whole Web site
- Classifications are the roots of hierarchies that organize large volumes of similar types of information
- Topic clusters are sets of related links
  - For example, sidebar and top of page *navigation* links
- *Content* information often fills the rest of a page

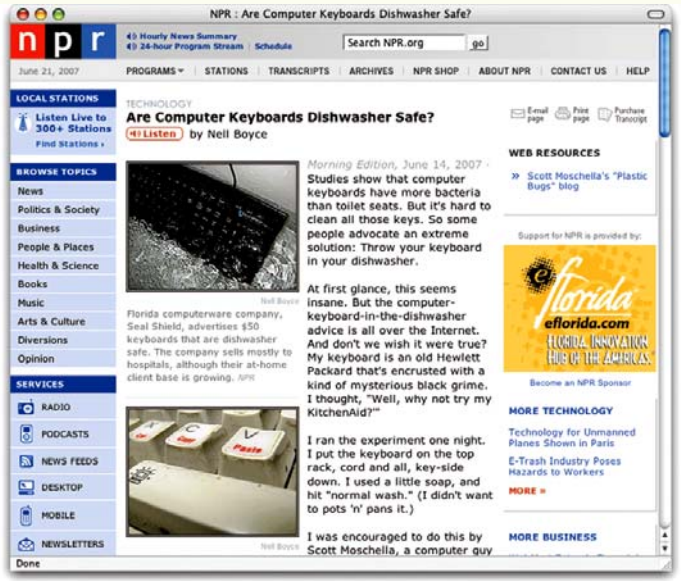


Figure 5.3 The National Public Radio (NPR) homepage http://www.npr.org.

Searching the Web for Information

- How a Search Engine Works
  - Two basic parts:
    1. Crawler: Visits sites on the Internet, discovering Web pages and building an *index* to the Web's content
    2. Query processor: Looks up user-submitted keywords in the index and reports back which Web pages the crawler has found containing those words
- Popular Search Engines: Google, Yahoo!, MSN, AOL, Ask

Crawlers

- When a crawler visits a website:
  - First identifies all the links to other Web pages on that page
  - Checks its records to see if it has visited those pages recently
  - If not, adds them to list of pages to be crawled
  - Records in an index the keywords used on a page (appear in the title, the body, or in anchor text)
- Crawlers can miss pages
  - No page points to it
  - Page is dynamically created on-the-fly
  - Page has only images
  - Page type is not recognized (not HTML, PDF, etc.)

Query Processors

- Gets keywords from user and looks them up in its *index*
- Even if a page has not yet been crawled, it might be reported because it is linked from a page that has been crawled, and the keywords appear in the anchor text on the crawled page
- Important to give the right terms to look up

Page Ranking

- Google's idea: PageRank
  - Orders links by relevance to user
  - Relevance is computed by counting the links to a page (the more pages link to a page, the more relevant that page must be)
    - Each page that links to another page is considered a "vote" for that page
    - Google also considers whether the "voting page" is itself highly ranked

Asking the Right Question

- Choosing the right terms and knowing how the search engine will use them
- Words or phrases?
  - Search engines generally consider each word separately
  - Ask for an *exact phrase* by placing quotations marks around it
    - "thai restaurants"

Logical Operators

- AND, OR, NOT
  - AND: Tells search engine to return only pages containing both terms (default)
    Thai AND restaurants
  - OR: Tell search engine to find pages containing either word, including pages where they both appear
    Thai OR Siam
  - NOT/-: Excludes pages with the given word
    -review
- AND and OR are *infix operators*; they go between the terms
- NOT/- is a *prefix operator*; it precedes the term to be excluded
- Google Help: Cheat Sheet
  - http://www.google.com/help/cheatsheet.html

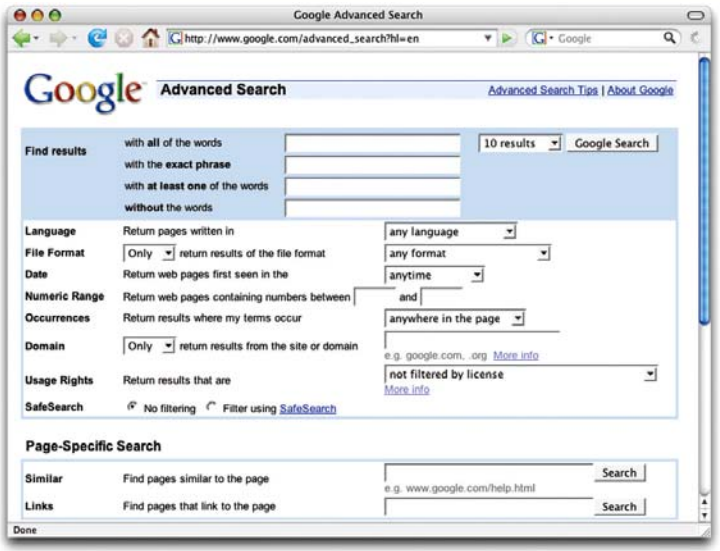


Figure 5.6 The Google search engine's advanced search view.

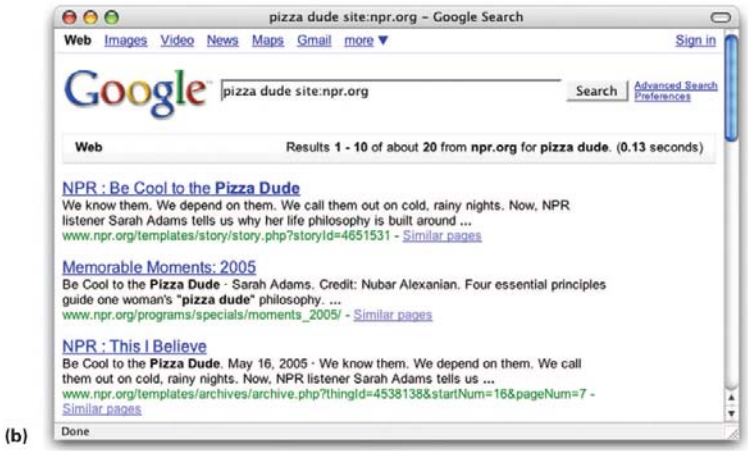
Five Tips for an Efficient Search

1. Be clear about what sort of page you seek (company or organization, reference page, etc.)
2. Think about what type of organization might publish the page you want
   - You might be able to guess the URL
3. List terms that are likely to appear on the pages you are looking for
4. Assess the results
   - Before looking at each returned page, check the results to see how effective your search was
5. Consider a two-pass strategy (focused searches)
   - Do a broad topic search, and then search within your results

Figure 5.7. Restricting the Thai restaurant "hits" by eliminating any page containing the word "vacation".

(a)

Figure 5.8 Focused searches for pizza AND dude: (a) NPR local search



(b)

Figure 5.8 Focused searches for pizza AND dude:
(b) Google search restricted to `npr.org`.

Web Information: Truth or Fiction?

- Anyone can publish anything on the web
  - Note prevalence of blogs and wikis
- Some of what gets published is false, misleading, deceptive, self-serving, slanderous, or disgusting
  - If it is on the web it must be true. – NOT!
- How do we know if the pages we find in our search are reliable?

Do Not Assume Too Much

- Registered domain names may be misleading or deliberate hoaxes
  - www.whitehouse.gov vs. www.whitehouse.org vs. www.whitehouse.com
- Look for who or what organization publishes the Web page
  - Respected organizations publish the best information available
- A two-step check for the site's publisher
  1. InterNIC (www.internic.net/whois.html) provides the name of the company that assigned the site's IP address, and a link to the WhoIs server maintained by that company
  2. Go to the WhoIs Server site and type the domain name or IP address again.
     - Information returned is the owner's name and physical address

Characteristics of Legitimate Sites

- Web sites are most believable if they have these features:
  - Physical Existence—Site provides a street address, phone number, e-mail address
  - Expertise—Site includes references, citations or credentials, related links
  - Clarity—Site is well organized, easy to use, and has site-searching facilities
  - Currency—Site was recently updated
  - Professionalism—Site's grammar, spelling, and punctuation are correct; all links work
- Remember that a site can have all these features and still not be legitimate. When in doubt, check it out (including cross checking). Ask a librarian.
  - Example: http://www.dhmo.org/ (Hoax about dangers of Dihydrogen monoxide – H_2O)

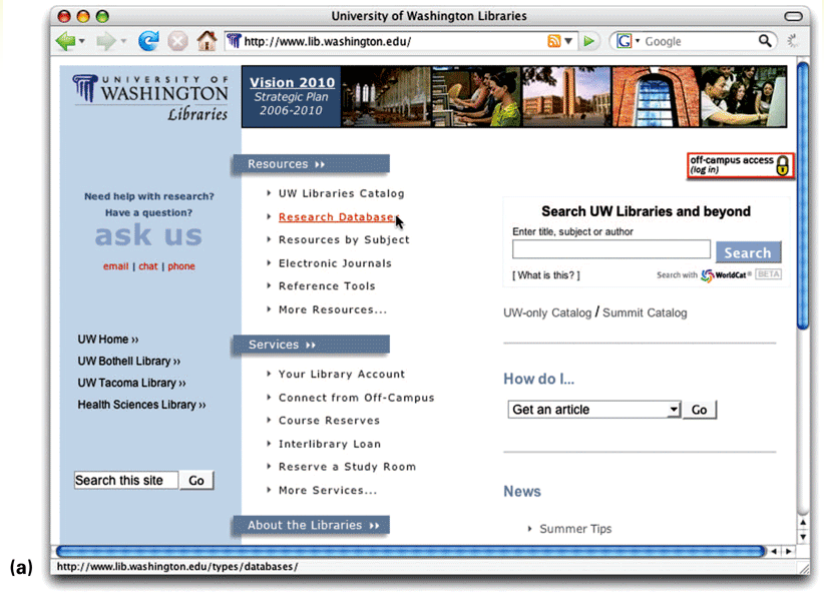


Figure 5.1 The (a) UW Libraries homepage,

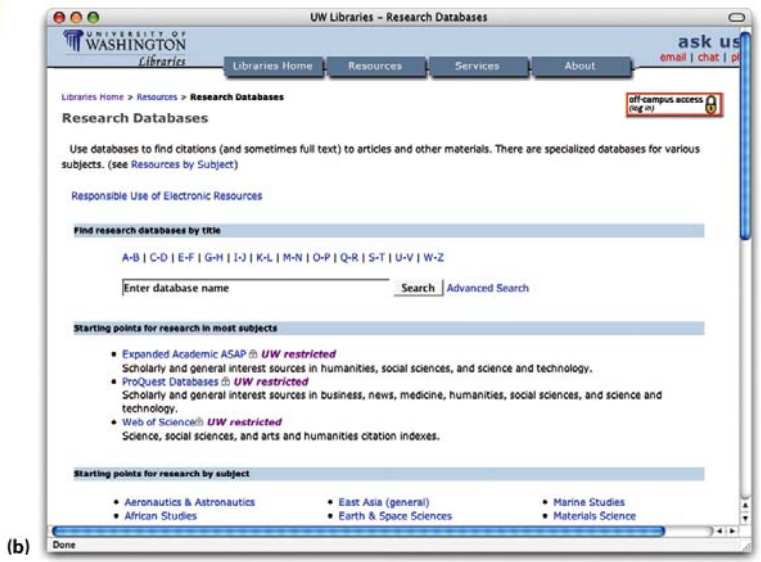
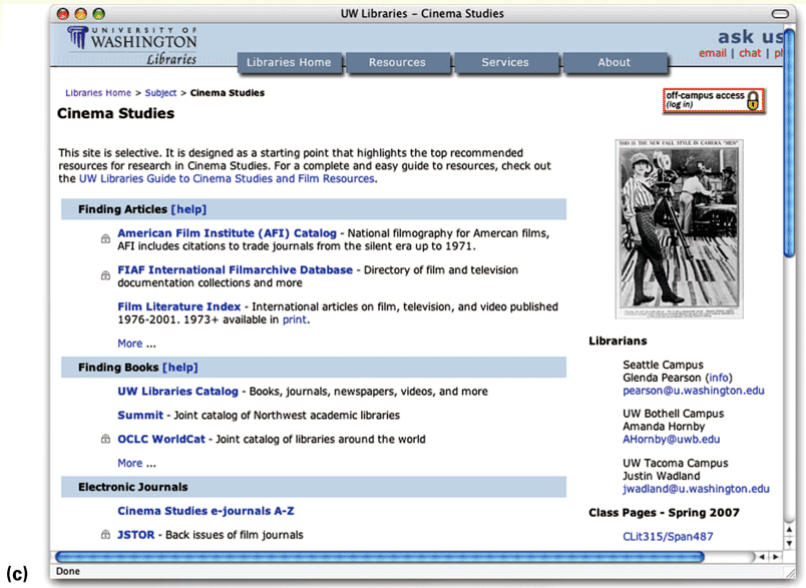


Figure 5.1 The (b) Research Databases page



(c)

Figure 5.1 (*continued*) The (c) Cinema Studies page.

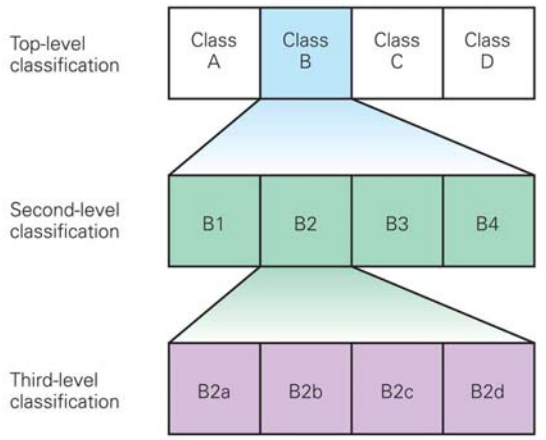


Figure 5.2. *Top-level, second-level, and third-level classifications of a collection.*

Table 5.1 The biological classification of human beings, *Homo sapiens*

Taxonomic Level	Name of Classification
Kingdom	Animalia
Phylum	Chordata
Subphylum	Vertebrata
Class	Mammalia
Order	Primates
Family	Hominoidea
Genus	*Homo*
Species	*sapiens*

(a)

Figure 5.4 NPR hierarchies: (a) Navigation links after clicking on *Music*.

(b)

Figure 5.4 NPR hierarchies: (b) *Archives* is the root of a huge hierarchy, with three ways of classifying the information.

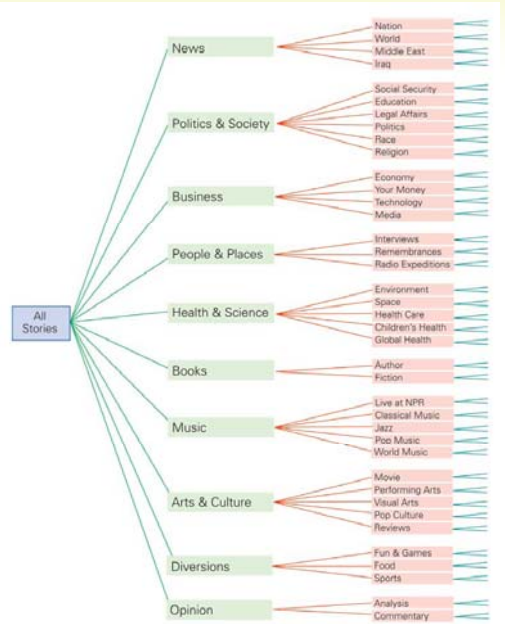


Figure 5.5. The NPR Story Hierarchy tree.

sailboard	220,000 hits
sailboard AND rentals	30,200 hits
sailboard AND rentals AND oregon	436 hits
sailboard AND rentals AND oregon AND "hood river"	360 hits
sailboard AND rentals AND oregon AND "hood river" –car	82 hits

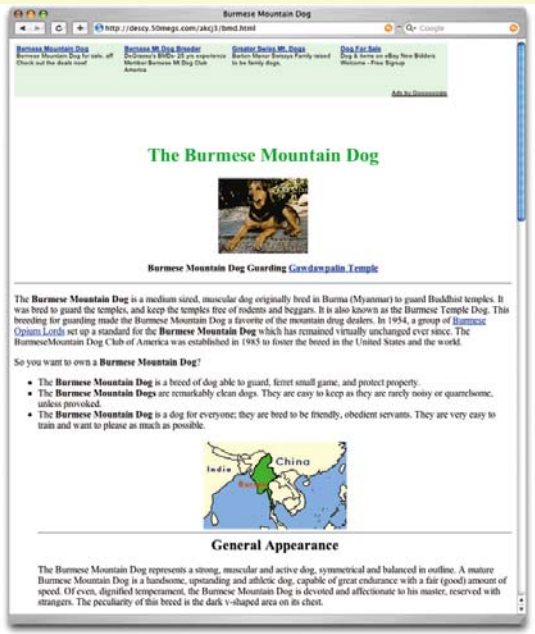


Figure 5.9 The Burmese mountain dog page.