

Search

# 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

# How Is Information Organized?

- Hierarchical classification
- 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

# Example

- NPR.org
  - [News](#)
    - [Nation](#) , [World Middle East](#) [Iraq](#)
  - [Politics & Society](#)
    - [Education](#) [Legal Affairs](#) [Politics](#) [Race](#) [Religion](#)
  - [Business](#)
    - [Economy](#) [Your Money](#) [Technology](#) [Media](#)
  - [People & Places](#)
    - [Interviews](#)
    - [Remembrances](#)
    - [Radio Expeditions](#)
  - [Health & Science](#)
  - [Books](#)
  - [Music](#)
  - [Arts & Culture](#)
  - [Diversions](#)
    - [Fun & Games](#)
    - [Food](#)
    - [Sports](#)
  - [Opinion](#)

# 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
    - "Going up in the hierarchy" means the classifications becomes more inclusive
    - "Going down in the hierarchy" means the classifications become more specific
    - 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
- Single links connect to very specialized pages

# 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 search, Alta Vista, Excite, InfoSeek

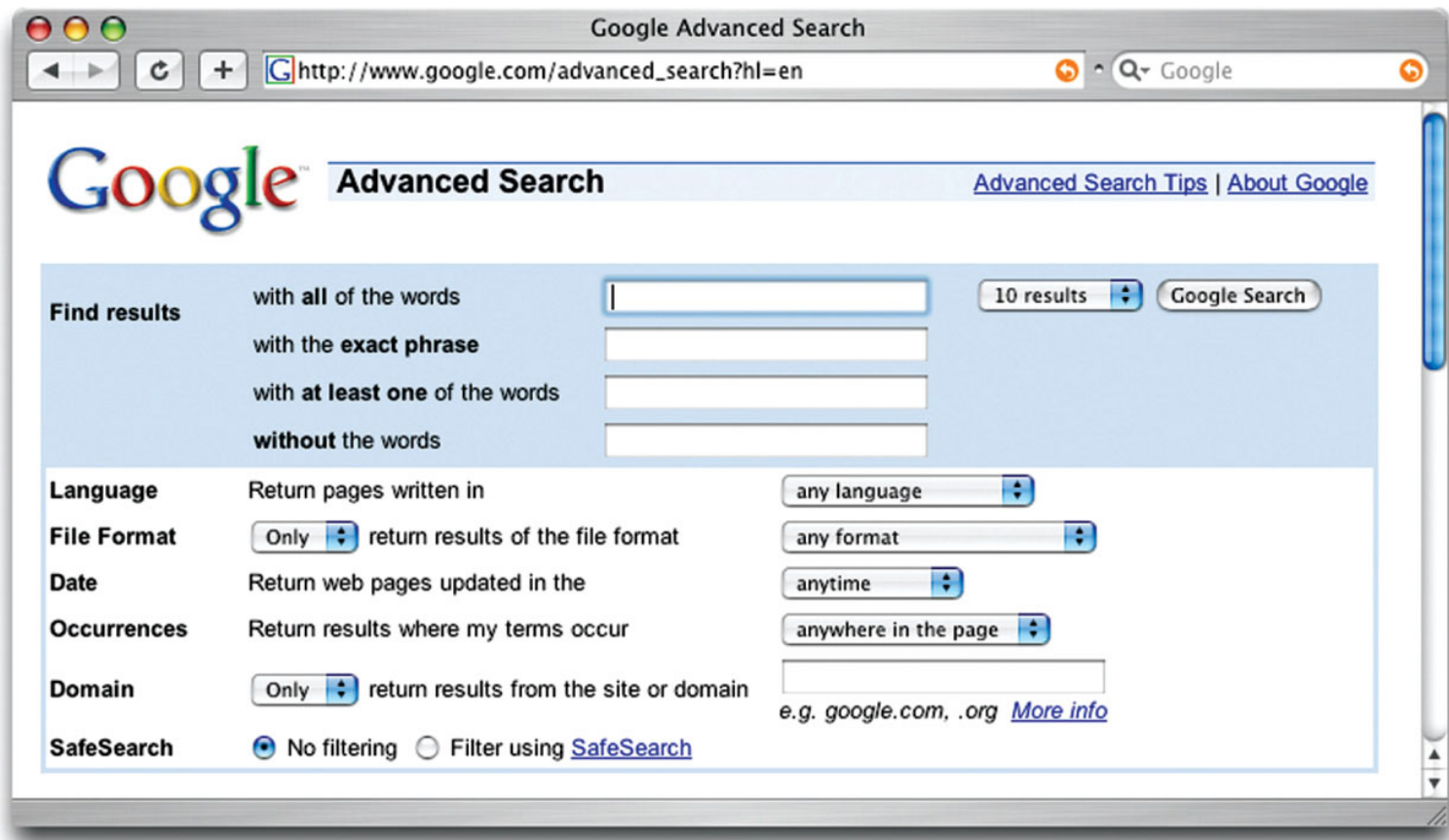


Figure 5.6. The Google search engine's advanced search view.

# 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

# Building the index

- Like the index listings in the back of a book:
  - for every word, the system must keep a list of the URLs that word appears in.
- Relevance ranking:
  - Location is important: e.g., in titles, subtitles, the body, meta tags, or in anchor text
  - Some ignore insignificant words (e.g., a, an), some try to be complete
  - Meta tag is important.
  - Page ranking



October 15, 2006

Programs and Schedules

Search NPR.org

find your local member station:

Call Letters find

e.g., "Austin, TX" or WXYZ or 20001

- News
- Politics & Society
- Business
- People & Places
- Health & Science
- Books
- Music
- Arts & Culture
- Diversions
- Opinion

POD NPR Podcasts

RSS News Feeds

Morning Edition

All Things Considered

Day to Day

Talk of the Nation

Fresh Air

DIGITAL CULTURE

# YouTube Founders Have the Last Laugh

Listen

Morning Edition, October 13, 2006 · Everyone seems to have something to say about Google buying YouTube. In fact, the video-sharing site's founders posted their own video about the purchase after it was announced.

E-mail this Page

Archives Transcripts Stations Shop About Contact Us Help

Copyright 2006 NPR Terms of Use Permissions Privacy Policy

E-mail page Print page Purchase Transcript

## MORE DIGITAL CULTURE

What Mom Doesn't Know About the Net

Internet as Archive, If You Have the Time

MORE >>

Support for NPR is provided by:



Become an NPR Sponsor

## MORE ARTS & CULTURE

'Scroll of Seduction' Tells Juana of Castile's Tale



```

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3c.org/TR/1999/REC-html401-19991224/loose.dtd">
<!-- saved from url=(0060)http://www.npr.org/templates/story/story.php?storyId=6259827 -->
<HTML xmlns="http://www.w3.org/1999/xhtml"><HEAD><TITLE>NPR : YouTube Founders Have the Last Laugh</TITLE>
<STYLE type=text/css media=screen>@import url( /templates/css/mainstyles.css );
</STYLE>

<STYLE type=text/css media=screen>@import url( /templates/css/bucket_alt.css );
</STYLE>

<STYLE type=text/css media="screen, print">@import url( /templates/css/stories.css );
</STYLE>
<LINK media=print
href="NPR YouTube Founders Have the Last Laugh_files/print_stories.css"
type=text/css rel=stylesheet><LINK title="Digital Culture"
href="/rss/rss.php?id=1049" type=application/rss+xml rel=alternate><LINK
title="Morning Edition" href="/rss/rss.php?id=3" type=application/rss+xml
rel=alternate>
<SCRIPT>
    if(navigator.userAgent.indexOf('PlayStation Portable') != -1)
    {
        document.write('<link rel="stylesheet" href="/templates/css/psp.css" type="text/css" />');
    }
</SCRIPT>

<META http-equiv=Content-type content="text/html; charset=iso-8859-1">
<META http-equiv=Content-Language content=en-us>
<META content=noarchive,index,follow name=robots>
<META content="Copyright (c) 2006 NPR" name=Copyright>
<META http-equiv=imagetoolbar content=no>
<META content=true name=MSSmartTagsPreventParsing>
<META
content="Everyone seems to have something to say about Google buying YouTube. In fact, the video-sharing site's founders pos
name=description>
<META content="" name=keywords>
<META content="" name=priority>
<META content=General name=Rating>
<META content="Living Document" name=doc-class><LINK href="/favicon.ico"
type=image/x-icon rel="Shortcut Icon">
<SCRIPT src="NPR YouTube Founders Have the Last Laugh_files/afwme.js"

```

# Page ranking

- PageRank is a weighted voting system:
  - a link from page A to page B as a vote, by page A, for page B
  - Weighted by the importance of the voting pages.
- You can see the page rank in the browser.
  - Try: [cnn.com](http://cnn.com), [yahoo.com](http://yahoo.com),
  - Compare: [ucdavis.edu](http://ucdavis.edu), [cs.ucdavis.edu](http://cs.ucdavis.edu)

# 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

# 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 to a page that has been crawled, and the keywords appear in the anchor on the crawled page

# 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

# Logical Operators

- AND, OR, NOT
  - AND: Tells search engine to return only pages containing both terms
    - Thai AND restaurants
  - OR: Tell search engine to find pages containing either word, including pages where they both appear
  - NOT: Excludes pages with the given word
- AND and OR are *infix operators*; they go between the terms
- NOT is a *prefix operator*, it precedes the term to be excluded

# Google search tips

- Case does not matter
  - Washington = WASHINGtOn
- Automatic “and”
  - Davis and restaurant = Davis restaurant
- Exclusion of common words
  - Where/how/and, etc. does not count
  - Enforce by + sign; e.g., star war episode +1
  - Enforce by phrase search; e.g., “star war episode I”
- Phrase search
  - “star war episode I”
- Exclude
  - Bass –music
- OR
  - Restaurant Davis Thai OR Chinese
- Domain
  - ECS15 site:ucdavis.edu
- Synonym search
  - ~food ~facts
- Numrange search
  - DVD player \$50..\$100
- Links:
  - <http://www.google.com/help/basics.html>
  - <http://www.google.com/help/refinesearch.html>

# Five Tips for an Efficient Search

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

# Web Information: Truth or Fiction?

- Anyone can publish anything on the web
- Some of what gets published is false, misleading, deceptive, self-serving, slanderous, or disgusting
- 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
- 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
  - InterNIC ([www.internic.net/whois.html](http://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
  - Go to the Whois site and type the domain name or IP address again.
    - Information returned is the owner's name and physical address

- -----
- Domain Name: UCDAVIS.EDU
- Registrant:
  - University of California at Davis
  - One Shields Avenue
  - Davis, CA 95616
  - UNITED STATES
- Administrative Contact:
  - Network Operations
  - University of California at Davis
  - One Shields Avenue
  - Davis, CA 95616
  - UNITED STATES
  - (530) 752-5999
  - netadmin@ucdavis.edu
- Technical Contact:
  - Network Operations
  - University of California at Davis
  - One Shields Avenue
  - Davis, CA 95616
  - UNITED STATES
  - (530) 752-5999
  - netadmin@ucdavis.edu
- Name Servers:
  - DNS-ONE.UCDAVIS.EDU 128.120.252.9
  - DNS-TWO.UCDAVIS.EDU 128.120.252.10
- Domain record activated: 19-Mar-1986
- Domain record last updated: 10-Feb-2000
- Domain expires: 31-Jul-2007

# Characteristics of Legitimate Sites

- Web sites are most believable if they have these features:
  - Physical Existence—Site provides a street address, phone number, email 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).

# Examples

- Google:
  - sailboat
  - Sailboat Sacramento
  - Sailboat Sacramento rental
- Search
  - Burmese Mountain dog (is this real?)
  - Bernese mountain dog