



Searching the WWW

Locating the right information on the WWW requires effort



Looking In the Right Place

The WWW is not the first place to look

- Go directly to a site -- www.irs.gov
- Guessing a site's URL is often very easy, making it a fast way to find information
- Go to right sight -- dictionary.cambridge.org
- Go to the library -- www.lib.washington.edu
- Go for the kind of information you want -- www.npr.org

Ask, "What site provides this information?"

2



Search Engines

No one controls what's published on the WWW ... it is totally decentralized
To find out, *search engines crawl* Web

- * Two parts
 - Crawler visits Web pages building an *index* of the content
 - Query processor checks user requests against the index, reports on known pages

Only a fraction of the Web's content is crawled



Google Advanced

Google Advanced Search

Find results with all of the words, with the exact phrase, with any of the words, without the words

Language: Return pages written in any language

File Format: Only return results of the file format any format

Date: Return web pages updated in the any time

Occurrences: Return results where my terms occur anywhere in the page

Domains: Only return results from the site or domain e.g. google.com, .org, .muse.edu

SafeSearch: No filtering, Filter using SafeSearch

4



Boolean Queries

Search Engine words are independent

Search for

- * Words don't have to occur together
- To be explicit about occurrences use Boolean queries and quotes
- * Logical Operators: AND, OR, NOT
 - monet AND water AND lilies
 - "van gogh" OR gauguin
 - vermeer AND girl AND NOT pearl

5



Queries

Searching strategies ...

- * Limit by top level domains or format
- * Find terms most specific to topic
- * Look elsewhere for key words, e.g. bio
- * Use exact phrase only when universal
- * If too many hits, requery
- * "Search within results" using "-"
- * Once found, ask if site is best source

6



Truth on the Web

- Much Web information is wrong
- Using the Web effectively means recognizing quality information
 - Information from reliable organizations is usually preferred -- check out ownership
 - Look for accuracy, currency, ...
 - Follow links to verify that the content supports the original page

Best: Locate independent verification

7



A Bogus Site

The Burmese Mountain Dog is a medium sized, muscular dog originally bred in Burma (Myanmar) to guard Buddhist temples. It was used to guard the temples, and keep the temples free of rodents and lizards. It is also known as the Burmese Temple Dog. In 1954, a group of Burmese, Chinese, and Thai set up a standard for the Burmese Mountain Dog which has remained virtually unchanged ever since. The Burmese Mountain Dog Club of America was established in 1983 to foster the breed in the United States and the world.

So you want to own a Burmese Mountain Dog?

- The Burmese Mountain Dog is a breed of dog able to guard, hunt small game, and protect property.
- The Burmese Mountain Dogs are remarkably clean dogs. They are easy to keep as they are rarely noisy or

www.50megs.com/akcj3/bmd.html

8



True Site, Bogus Implication

DHMO.org
Dihydrogen Monoxide Research Division

WELCOME

Welcome to the web site for the Dihydrogen Monoxide Research Division (DHMO), currently located in Newark, Delaware. The controversy surrounding dihydrogen monoxide has never been more widely debated, and the goal of this site is to provide an unbiased data clearinghouse and a forum for public discussion.

Explore our many Special Reports, including the DHMO FAQ, a definitive primer on the subject, plus reports on the environmental consequences, current research, and an insider's view of the industry.

www.dhmo.org

9



Intellectual Property

Most intellectual property (IP) is protected

- * You can't use it unless you pay the creator
- * IP: movies, songs, performances, photos, Web pages, sculptures, ...
- * Penalties are severe ...

You can't publish stuff off Web, e.g. photos, w/o authorization – pub domain, allowed, permission



Copyright

Applies to writings, photos, programs, ...

- * No © notice is required
- * More rights than copying
- * Noncommercial use is no excuse
- * Penalties are huge: \$100,000 each
- * Fair use is for worthy uses (education)
- * See Chapter 12, pp.353-358

Bottom line: Use your own intellect to create your own intellectual property ... that way you're paid



Page Rank

Millions of hits make no difference if the one you want is buried in the list

- * Google solves this using page rank
- * A page's rank is based on the number of pages that reference it and their rank
- * Page rank is Google's measure of importance
- * Pages are listed in decreasing rank

12



Google Whacking

Google Whacking is a game for people with no social life ...

- * Find a pair of words which have only one Google hit
- * Search googlwhack for lists
- * If you list your GW on your Web page, guess what ...!

ambidextrous scallywags
illuminatus ombudsman
squirreling dervishes
assonant octosyllable
fetishized armadillo
panfish interrogation

13



Google Bomb

Google's page rank can be affected by users ... the Google Bomb

- * Many users using a common term to link to a site (term force it to be listed first in a Google search ...
- * "miserable failure" has George W Bush's biography as the first hit

Sponsored Links
[Biography of President George W. Bush](#)
Biography of the 43rd President of the United States.
[www.whitehouse.gov/president/gwbio.html](#) - 25k - Cached - Similar pages

[Why these results?](#)
These results may seem politically slanted. Here's what happened.
[www.google.com/googleblog](#)