Searching for Information

INFO/CSE 100, Autumn 2004
Fluency in Information Technology

http://www.cs.washington.edu/100

Readings and References

- Reading
  - *Fluency with Information Technology*
    - Chapter 5, Searching for Truth

- References
  - Research 101, UW Libraries
    - *Research 101* is intended to help students learn to perform basic information research so they can tackle information problems anywhere.
    - http://www.lib.washington.edu/uwill/research101/

Searching for Information

- The Web and its search engines have made it much easier to quickly find the top-level information about a topic
  - A simple Google or Yahoo search will turn up *something* on almost any topic
  - That's enough if you're just trying to get a general idea about a topic
- But what if you really care about the accuracy?
  - Look deeper, cross check, work with experts

Information hierarchies

- Experts in a particular field usually organize the information about the field into a structure
  - If you understand the overall structure and ...
  - you want some detail, then ...
  - you can move quickly through the structure to the particular item of information that you seek
- This works very well
  - IF you know the structure well enough to navigate within it.
Library Classification Systems

- Within a library collection, materials are typically organized by subject. Librarians assign a call number based on a work's subject
  - Library of Congress
    - Used in most college, university, and research libraries because it handles large collections
  - Dewey Decimal
    - Used in most public and school libraries because it is more effective for smaller collections

Library of Congress Classification

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>Plantae</td>
<td>Organisms that usually have rigid cell walls and usually possess chlorophyll.</td>
</tr>
<tr>
<td>Subkingdom</td>
<td>Embryophyta</td>
<td>Plants forming embryos.</td>
</tr>
<tr>
<td>Phylum</td>
<td>Tracheophyta</td>
<td>Vascular plants.</td>
</tr>
<tr>
<td>Subphylum</td>
<td>Pterophyta</td>
<td>Generally large, conspicuous leaves, complex vascular system.</td>
</tr>
<tr>
<td>Class</td>
<td>Angiospermae</td>
<td>Flowering plants, seed enclosed in ovary.</td>
</tr>
<tr>
<td>Subclass</td>
<td>Dicotyledoneae</td>
<td>Embryos with two seed leaves.</td>
</tr>
<tr>
<td>Order</td>
<td>Sapindales</td>
<td>Sapindaceae order consisting of a number of trees and shrubs.</td>
</tr>
<tr>
<td>Family</td>
<td>Aceraceae</td>
<td>Maple family.</td>
</tr>
<tr>
<td>Genus</td>
<td>Acer</td>
<td>Maples and box elder.</td>
</tr>
<tr>
<td>Species</td>
<td>Acer rubrum</td>
<td>Red maple.</td>
</tr>
</tbody>
</table>

From PhysicalGeography.net, http://www.physicalgeography.net/fundamentals/9b.html

Press 1 for sales, press 2 for ...

• Hello, thank you for calling ...  
  » press 1 for sales  
  » press 2 for frequently asked questions  
  » press 3 for account information  
  » press 4 customer support
• Poorly designed phone menu hierarchies are a royal pain  
  » dead ends are a waste of time  
  » multiple steps to get to the desired information  
  » designed to reduce costly customer interaction

Is the Web a hierarchy?

• Not a hierarchy, and that's good (to some)  
  » A major part of its amazing success is that relevant information is just a click away,  
  » That's why it's "The Web" not "The Tree"
• Not a hierarchy, and that's bad (to some)  
  » Any web page can link to any other web page  
  » This is causing heartburn among the gatekeepers  
  » You can link directly to a page on a web site without going through a top-level page

In the beginning

• When the Web was new  
  » there was no hierarchy and ...  
  » there was no way to find what you wanted other than careful detective work and keeping good notes
• Web detective games  
  » How quickly can you find the following factoid?  
  » Quite a challenge - those who knew their links could find information quickly

Now: Scan, then drill down

• Search engines do the top-level scanning for us  
  » Enter your search term in Google, Yahoo, etc  
  » You will get back a list of sites that are in some way related to your topic
• The links are entry points into the giant hairball that is the web  
  » It's still not a hierarchy, but you can get close to the high value sites quickly  
  » Drill down for information within those sites
Now: Select, then drill down

- Search engines are not the only way to get to the top of an information rich tree
  - UW Administrative information?
    - UW Information Navigator page
  - Scholarly information?
    - UW library has high-density information
  - Tax forms and laws?
    - The IRS web site is the definitive source

Search Engines

- No one controls or assigns hierarchy locations to the pages published on the WWW ... it is totally decentralized
- To find pages, 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

Search Engine words are independent
  - The words don’t have to occur together on the page
- 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

See also http://www.google.com/help/refineSearch.html
Queries

- Searching strategies …
  - Limit by top level domains or format
  - Find terms most specific to topic
  - Look elsewhere for key words
  - Use exact phrase only when universal
  - If too many hits, requery
  - Once found, ask if site is best source

Selected Sites

- For some searches, there are information rich areas in which to start. No need to Google.
  - Scholarly material
    - UW Library
    - Library of Congress
  - Government information
    - http://access.wa.gov/
    - http://www.firstgov.gov/
The Web and reality

- The Web contains only a fraction of the information available to us
  - It provides a great tool for getting started with a topic
  - It provides rapid access to many significant collections of information
- The real world contains much, much more!
  - Use the Web as a map, then go exploring
  - Go to the library, talk to a librarian, read the primary sources, talk to the experts in the field