FIT 100

Searching for the Right Stuff

And boy, is there a lot of stuff to search through!

The best way to narrow down the amount of stuff you have to look through is by searching in the right area to begin with

FIT 100

Finding and Evaluating Information

- Key principle:
  - The best place to search for something is where it is likely to be found.
  - Question: Where is the best place to find tax information?
    - Used car information?
- What are the potential places that one would find the best information?
  - Library – not everything is on the Internet, and it never will be! – and the library specializes in organizing information.
  - Specialized databases – if you do search for information electronically, go to the places where the content has already been evaluated and authenticated. Often these electronic databases can be used, for free, at ….the Library!
- So why all this focus on the library?

FIT 100

What do Humans do?

- We associate things with other things and we organize them (group them together)
- Humans have a natural tendency to organize (cluster) similar things together
  - When you were a child, did you put all of your Barbie dolls in one place, or all of your robot toys? How about the same color toys?
- Think about the subjects you study in school: Math, Social Sciences, Geography, Art, etc.
  - Each of these subjects is further divided, becoming more and more detailed, specialized
- Librarians tackle the problem of trying to organize and place information where people can most easily find it

FIT 100

Organizing Electronic Things

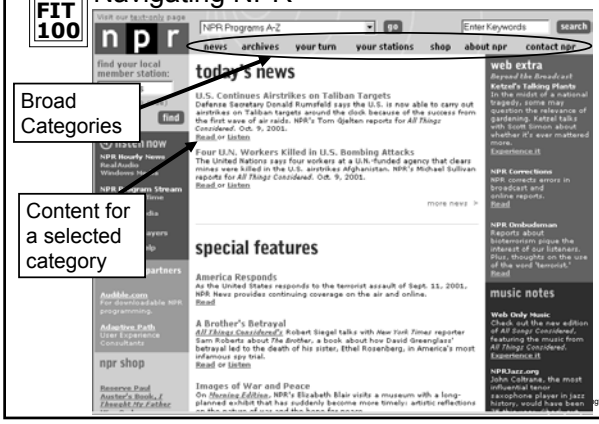
- If we tend to group similar things together, then we also tend to look for things that are similar in the same places
  - Problem: Do any 2 people organize things in exactly the same way? Do any 2 groups of people?
- Now, go to the web where all of that stuff sits
  - NO organization
- But, individual web sites can try to organize their “stuff”. A well-designed and well-organized site will help you find the “stuff” faster

FIT 100 A Well Organized Site

- A web site that wants to help users find relevant information fast is usually organized in a manner similar to a Library:
  - The site is divided into broad categories (subjects)
  - Each category in turn has sub categories
    - Familiar? Didn't we try to organize folders on Dante in a meaningful manner?
- If a user that visits a site (you) can quickly see how things are categorized then they will find what they need much quicker
  - This assumes they have started looking in the right spot…

http://www.lib.washington.edu

FIT 100 Navigating NPR

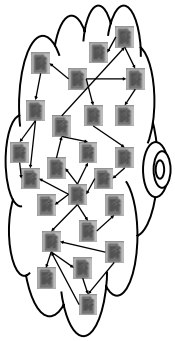


FIT 100

But what if you JUST DON'T KNOW?

Then the next step may be to go to do a search of information on the Web

FIT 100 Structure of the Web



- The Web is just lots of documents on lots of machines
- Each document has a link to it from some other document
- If Page A links to Page B, you can get to page B from A (but not vice versa!)
- However, searching for information by going from page to page, one document at a time, would take the rest of eternity

FIT 100 Directories: Yahoo!

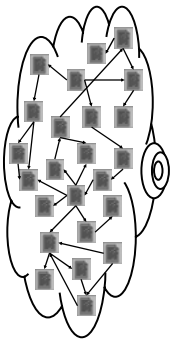
- Not a search engine!
- A directory
  - Web pages have been organized into a hierarchical structure based on broad categories
  - Humans do the organizing, usually

Arts & Humanities Literature, Photography...	**News & Media** Full Coverage, Newspapers, TV...
Business & Economy B2B, Finance, Shopping, Jobs...	**Recreation & Sports** Sports, Travel, Autos, Outdoors...
Computers & Internet Internet, WWW, Software, Games...	**Reference** Libraries, Dictionaries, Quotations...
Education College and University, K-12...	**Regional** Countries, Regions, US States...
Entertainment Picks, Movies, Humor, Music...	**Science** Animals, Astronomy, Engineering...
Government Elections, Military, Law, Taxes...	**Social Science** Archaeology, Economics, Languages
Health Medicine, Diseases, Drugs, Fitness...	**Society & Culture** People, Environment, Religion...

FIT 100 What is a Search Engine?

- A collection of programs designed to assist users in finding information
- Consists of four things (the book gives the 2 main ones):
  - A crawler (aka spider, robot)
  - A query processor
  - A user interface
  - A database
- A crawler does what the name implies:
  - "Crawls the Internet" building an index of URLs and key terms that are hopefully an indicator of the content of the page.
- A query processor takes a request from the user (search terms)
  - Retrieves the list of URLs associated with a given set of key word terms according to the index

FIT 100 A Web Spider: See how they crawl!



- Start at a document
- Repeat:
  - Store document
  - Extract all links
  - Index Terms
  - Follow every link
- Until: all links followed
- Redo when necessary
- Need to pick "top" page, and good pages
  - Pages with multiple links

FIT 100 The Search, literally, never ends

- Crawling is an ongoing process. Why?
  - To keep up with the millions of new pages that are added the the Web on a weekly basis
  - To go back and revisit sites and make sure that links aren't broken
- Lists of URL's are created on the spot for users
  - But the query processor doesn't have a robot search the web each time
  - Goes to the database of searched pages and matches search terms with index terms
- The effectiveness of a search engine will depend in part on how much of the Internet the crawlers have seen and how discriminating the index is.

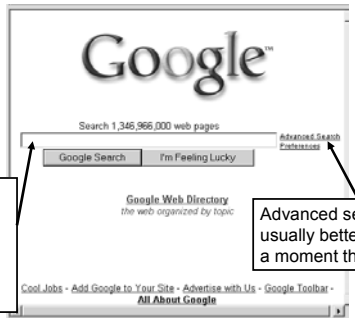
FIT 100 Search Technique

- Search engines use basic logic to determine how to answer your question
- Boolean logic, in the form of Boolean operators, are the foundation of search logic:
  - AND
  - OR
  - NOT
- Many search engines now use "search math" instead of boolean terms
  - +, -
- Search for exact phrases, like titles, with quotation marks
  - "Fellowship of the Ring"

FIT 100 Search Technique

- Successful search takes more than just Boolean logic
  - There is too much out there
- Other search mechanisms
  - Ranking of search terms
  - Weighted terms
  - Proximity
- Some search engines have incorporated other ways to return better results

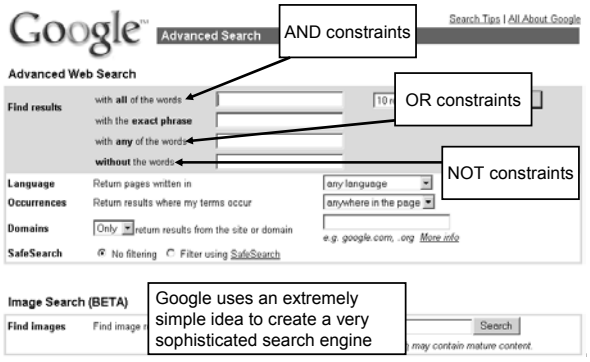
FIT 100 A bit about Google



The "basic" search, seen here, is often effective just based on their rankings

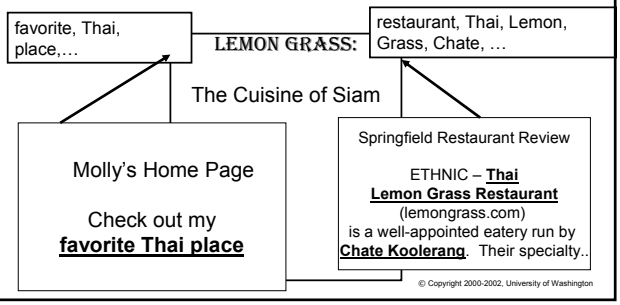
Advanced search is usually better if you spend a moment thinking about it

FIT 100 Advanced Search



FIT 100 Google's Sophistication

- Google is a search engine that indexes a page by the words used in the anchor tags of pages that link to it



FIT 100 More Google Sophistication

- Popularity is also a key to Google's rankings
- If page A links to page B, then that is considered a vote by page A for page B
- If page A is also a very popular site that many other sites link to, then page A's vote is worth more
- How can WE get a sense of how popular a web site is? Do a simple link search in Google to see how many sites link to the site that interests us.

FIT 100 Summary

- Search where you are most likely to find the information
- Good sites will have effective navigation that you can easily figure out
- Local searches in good sites can quickly find candidate pages
- Search Engines build indexes to assist in searching the web
- When doing a search of two or more words or phrases, specify whether
  - Both words MUST be present: AND, +
  - At least one of the words must be present: OR
  - The word(s) must NOT be present: NOT, -