Chapter 2: Searching for Truth: Locating Information on the WWW

Virtually every major uses the internet to help with research.
Virtually every business uses the internet to help with communications, marketing, sales.
To be better users of the internet, means we will be more productive and more valuable to our employers!
Make your knowledge of computers visible on your resume! It will set you above the crowd and get you noticed.

This chapter talks about just one kind of skill – searching. Computer skills are mainly “self taught”. We are not in Kansas anymore, Toto!

By working in groups, you have likely discovered that you learn a lot by watching others use the computer.

Often you don’t even know what skills you are missing. Be alert for things you should know!

Fake it until you make it! Ask questions. Describe your problems accurately. Make lists of what you want to know.

Searching in All the Right Places

Often we “google it” – but don’t really note what website we are at.
• 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

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
Today's lecture

• Need you to shut your laptops. Too distracting.
  Embarrassment/penalties are not my style – but will if needed.
• Transition – work in groups, listen to professor. Needs to be quick.
• A intelligence, B attitude
• Stay focused. If you are only coming to class to get credit for the
  inclass assignment, you are NOT getting your money’s worth!
  
• If you struggle: Read ahead – both text and notes. Daughter’s
  complaint about not understanding international professor.
• Homework is a mixture of in-class and out of class work.

Today’s lecture

Today we will talk about one of the great ideas of computer
  science.
Computer science is not so much about computers as it is
  about problem solving! Very creative!
Today we will talk about a great solution to a problem

People at USU

Interesting people together

Faculty

Students

married

Single

USU

Boring

Interesting

Average

Tall

married

Single

People at USU

All single people – easy to find.
Want to invite all interesting people to a party, not easy to find.
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

As computer scientists, we worry about information storage and retrieval. If things are stored poorly, retrieval is slow.

If we have to sort through a lot of “stuff” we aren’t interested in, we lose. Organization is the key.

Other choices – Network/graph Model

Preventative Maintenance

Rigid

Flexible

Spall

Joint

Crack Seal

Patch

Silicone

Asphalt

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

preventative maintenance>rigid>spall
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) — e.g. townships in Iowa, or be partitioned (every category appears only once) Cities/Counties in Utah.
- Number of levels may differ by category, even in the same hierarchical tree. Can you think of an example where this is true?

Organization concept – a tree graph looks a bit like a tree shown upside down compared with a real tree

- Internet: Yahoo! subject index
- Information management: Dewey Decimal System
- Management: hierarchical organizational structures
- Computer Science: search tree
- Biology: evolutionary tree
- Business: pyramid selling scheme
- Project management: work breakdown structure
- Linguistics (syntax): Phrase structure trees

Dewey Decimal System

- 000 – Computer science, information, and general works
- 100 – Philosophy and psychology
- 200 – Religion
- 300 – Social sciences
- 400 – Languages
- 500 – Science and Mathematics
- 600 – Technology and applied science
- 700 – Arts and recreation
- 800 – Literature
- 900 – History and geography and biography
- So where would you put Java — it’s a language for computers?
- What about the biography of Henry Eyring — the chemist — it is biography or Science and mathematics?

Hierarchy doesn’t always make sense – “Graph” structure
Hierarchy doesn’t always make sense in company organization

Visit with your neighbor
- So how is a website organized?
  Hierarchically (tree-like) or Graph-like?

Cite evidence to prove your point!
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 links or top of page navigation links
- Content information often fills the rest of a page

Great idea of computer science

Search Engine (such as google)

- What problem were they trying to solve?

- How does a search engine work?

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. Suppose you wanted the webpage you created to be seen by a crawler, what would you do? Test it. Try putting some unique keywords on your webpage and see if you can find it via google? Google your own name.
    2. Query processor: Looks up user-submitted keywords in the index and reports back which Web pages the crawler has found containing those words. Have you ever wondered why the computer is so fast to find what you are looking for when there are billions of pages? (19.2 billion in 2005 indexed by Yahoo)
- 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)

**Index Words**

- Unless author of webpage specifies the keywords (meta tag), the search engine must determine them.

- Need an automated way of determine which words are important
  - words in title
  - words near beginning (layout subject matter)
  - words repeated several times
  - ignore common or stop words: life is a dream

  Shows you what is ignored:

  Results 1 - 10 of about 48,300,000 for life is a dream. (0.30 seconds)

**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.)

**Index Words**

- Problems with keyword searching?
  - Spelled same, but mean something different: hard cider, hard stone, hard exam, computer hard drive
  - stemming – big, bigger, biggest
  - Plurals, verb tenses

**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
But how does it figure out what I want? Page must have the word I’m looking for – but which to give me?

Page Ranking

- Have you any experience with a BAD search engine? What happens?

Google's idea: PageRank

- FAVORS pages with your words in PHRASES, CLOSE TOGETHER, and IN THE ORDER TYPED
- 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

Key to success: 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"
  - Ignore connecting words

Testimonials

- What do you do with google search?
Search Basics: Searching made easy

- **Tip #1: Choose your keywords wisely.** Select descriptive, specific words.
  - Search for: Vancouver 2010 Olympics rather than Olympics
  - Search for: 2004 football statistics rather than football

- **Tip #2: Use quotes for an exact match.**
  - When you enclose your search query in quotation marks, you’ll only get results for the exact terms you entered in the order you entered them.
  - Search for: “There are strange things done” rather than There are strange things done

- **Tip #3: Try one of our search features.** When you search for a number, it’s not necessarily going to do its normal web search, looking for pages with that number in them.
  - For calculations, search for the equation. Example: 5+2*2
  - For flight status, search for the flight number. Example: united 134
  - To track a package, search for the package’s tracking number. Example: 1Z9999W99999999999

Choosing Keywords

- **Synonym Search**
  - If you want to search not only for your search term but also for its synonyms, place the tilde sign (~) immediately in front of your search term. ~food matches recipes, nutrition, cooking

- **Dictionary Definitions**
  - To see a definition for a word or phrase, simply type the word “define” followed by the word or phrase. Note that the results will define the entire phrase.

Things to try – Why handled differently?

- **Weather**
  - To see the weather for many U.S. and worldwide cities, type “weather” followed by the city and state, U.S. zip code, or city and country.

- **Stock Quotes**
  - To see current market data for a given company or fund, type the ticker symbol (BAC MSFT JPM AAPL) into the search box. On the results page, you can click the link to see more data from Google Finance.

- **Time**
  - To see the time in many cities around the world, type in “time” and the name of the city.

- **Sports Scores**
  - To see scores and schedules for sports teams type the team name or league name into the search box.

- **Calculator**
  - To use Google’s built-in calculator function, simply enter the calculation you’d like done into the search box.

How does Google’s “Did you mean” feature work?

- Google’s spell checking software automatically looks at your query and checks to see if you are using the most common version of a word’s spelling.
- If it calculates that you’re likely to generate a greater number of search results with an alternative spelling, it will ask “Did you mean: (more common spelling)?”
- It is completely automated, so you can’t change when it is invoked.
- Some businesses were concerned that when people typed in their business name, google asked “Did you mean” competitors name. Sorry. There is nothing you can do about that – just get your name out there.
  - Example: Hope Depot
- Because Google’s spell check is based on occurrences of all words on the Internet, it is able to suggest common spellings for proper nouns (names and places) that might not appear in a standard spell check program or dictionary.
Logical Operators

• Can use “advanced search” to restrict searching.
• AND, OR, NOT (Requires capitalized AND, OR, NOT)
  – AND: Tells search engine to return only pages containing both terms (default).
  – OR: Tells 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 Help: Cheat Sheet

Boolean (true/false) operators

• Logan AND (Vicki OR Allan)
• (Logan OR Faculty) AND (Vicki OR Allan)
• OR: one or the other or both
• XOR (not available in this context): exclusive or. One or the other but not both. You can get overtime pay or earn time off.
• We use parentheses to group.
• Look at number of results to tell if your search operators are working.

• stalker +in +the group
• Eliminates pages not containing in and the somewhere.
Useful Search Operators (restrict search to terms occurring in specific parts of the web page)

- **site:** Search only one website. `site:www.usu.edu` (Search USU site for club info.) Great – when local search engine is BAD.
- **inurl:** require terms to be in the url somewhere
- **allintitle:** restrict results to those with all of the query words in the title
- **intitle:** For instance, `[intitle:google search]` will return documents that mention the word “google” in their title, and mention the word “search” anywhere in the document (title or no). Note there can be no space between the “intitle:” and the following word.

Useful Search Operators (from basic search window)

- `[#]...[#]` Search within a range of numbers. `DVD player $100..150` (Search for DVD players between $100 and $150)
- **link:** linked pages. `link:www.cs.usu.edu` (Find pages that link to the USU computer science website.)
- **info:** Info about a page. `info:www.usu.edu` (Find information about the USU website.)
- **related:** Related pages. `related:www.usu.edu` (Find websites related to the USU website.)
- **filetype:** limit type of document. `filetype:ppt site:edu china one-child policy` `filetype:xls OR filetype:pdf death-rate malignant mesothelioma "united states"

Locate dictionary definitions and thesaurus terms (synonyms, antonyms, etc.) using the link in the blue statistics bar after you search:

- scholarship
  
  Results 1 - 10 of about 47,000,000 for scholarship [definition]

- Click Translate the page after a page in a foreign language
  
  1. Go to Google Advanced Search
  2. Limit to PAGES IN FRENCH and search jacques chirac avec george bush

IN Class Assignment (groups of 2,3,or 4) submit to Blackboard

For each of these search tasks, write a few sentences describing your search steps: what you tried searching on, number of hits, and relevant pages you found.

1. Try out a search with **AND, NOT, OR.** Explain how you know the search is working.
2. What time is it in Copenhagen?
3. Use **title:** to look for all sites regarding global warming.
4. Try three other searches using special features described today.
5. You want to find documentation about how technology training will make you more hireable.
6. Using the site: option, look for stories on biofuels from NPR.
7. You are looking for the author of the saying that goes SOMETHING like: If you haven’t make a mistake, you have never tried anything new.
• Stemming: Google now uses stemming technology. Thus, when appropriate, it will search not only for your search terms, but also for words that are similar to some or all of those terms. If you search for **pet lemur dietary needs**, Google will also search for **pet lemur diet needs**, and other related variations of your terms. Any variants of your terms that were searched for will be highlighted in the snippet of text accompanying each result.

• **I’m feeling lucky** - Takes you directly to first web page returned for your query. Hit this button instead of the regular “google search” button.

---

**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
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
  - (some use bad language and state they are not official)
  - You BUY the domain name – so whoever asks first, gets it.
- 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
  - didn’t always work for me.

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
  - Current—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 – H₂O water)
    - Odd sense of humor? Eaton article in Herald Journal 1/14/09 – teach Jim McMahon to throw a football, wife’s cocaine habit.
    - Destructive – anti-X propaganda

Try this yourself – as the numbers have changed

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sailboard</td>
<td>220,600 hits</td>
</tr>
<tr>
<td>sailboard AND rentals</td>
<td>26,260 hits</td>
</tr>
<tr>
<td>sailboard AND rentals AND oregon</td>
<td>436 hits</td>
</tr>
<tr>
<td>sailboard AND rentals AND oregon AND “hood river”</td>
<td>380 hits</td>
</tr>
<tr>
<td>sailboard AND rentals AND oregon AND “hood river” –car</td>
<td>62 hits</td>
</tr>
</tbody>
</table>
InClass Assignment

- Using Advanced Search with google, update the numbers found on the previous slide.
- For each of your majors (pick one you are thinking about if you don’t have one), find the expected starting salary and the percent of people who graduate in that major who find a job which is relevant to their degree.