

FIT 100
Assignment 1: Searching the Web
(or, Finding what you want, and no more!)

Winter 2002

Due: Friday, January 18, beginning of lecture. No late assignments accepted.

Format: Write your answers using any computer writing tool (text editor, word processor, etc.). You don't need to include the original questions, just the question numbers. Be sure your name and section are at the top (points off if not!). Print. Staple all pages together (PS. Don't count on there being a stapler in lecture!). Bring to lecture.

Required reading for Assignment 1:

Link to and read the sections on **Search Engine Math** and **Boolean Searching** at the Search Engine Watch website. Review the **Search Engine Features** page to help in your search:

Search Engine Math

<http://www.searchenginewatch.com/facts/math.html>

Boolean Searching:

<http://www.searchenginewatch.com/facts/boolean.html>

Search Engine Features for Searchers:

<http://www.searchenginewatch.com/facts/ataglance.html>

Introduction:

Many of you have done a fair amount of browsing and searching on the Internet. But have you ever thought about how and where to search in such a way that you get only those sites you want and no more? Constructing a search that does exactly that is very difficult, if not impossible. However, you can learn to search the Web in a way that brings back a smaller set of "hits" (web pages that match your search), and improve the chances that these hits are more relevant than not.

So, what exactly IS a Search Engine? And why do I care?

A search engine is really just a program, or series of programs, that is designed to try and help users find useful information on the Web. A search engine consists of several pieces (these will be covered in lecture). The basic idea is that a search engine takes terms that you enter and tries to match those terms with documents out on the Web that are most relevant.

Seems simple, doesn't it? Yes, it seems simple... but relevance is hard for a program to determine when it doesn't "know" the person doing the search. This is an exercise for you to see both the ease and difficulty of searching for information on the web.

Objectives:

- To use basic search strategies in a search engine and bring back sites with information on a topic.
- Learn to find the best search method for a particular search engine.
- To develop systematic and precise search skills.

Online Resources:

Some available search engines (but not the only ones!!!!):

Google: <http://www.google.com/>

Uses link popularity as a way to rank a web site. If 50 different sites link to one other site, this is a good indicator that it is a relevant page for the topic it covers.

AltaVista: <http://av.com/>

One of the largest search engines around. Allows searches just on images and other formats. Also has a translate feature.

DogPile: <http://www.dogpile.com/>

DogPile is a metasearch engine. It runs a search across other search engines to get results. It allows you to specify a search for images or audio files, etc.

Some search engines use a directory structure to organize web sites by subject:

Yahoo!: <http://www.yahoo.com/>

Directory setup. Provides email, news, etc.

List of Search Engines by function:

<http://www.searchenginewatch.com/links/>

A useful page to go to lists of search engines.

To Do:

1. Go to Yahoo.com and use the categories to find the web site for the Computer Science Department at the University of Washington. What is the most logical starting point?

After you have found the CSE site, then go back to the start page at Yahoo and try to search for the same thing using the search box at the top of the page. How did you search? Did the UW site come up in the first page of results?

2. Search for information about the disturbances in Seattle in December of 1999 over WTO, the World Trade Organization.

How did you construct your search?

Compare several search strategies. Which one appears to be more effective? (look at your top 10 results)

Can you figure out what is happening as the results are returned? Are pages being brought back because they have all of the terms? Or because they have just some of the terms?

3. Find a site dedicated to the victims of the terrorist attack in September and give the URL. How did you construct your search?

4. Using the list of search engines by function at:
<http://www.searchenginewatch.com/links/>

What would be a good engine to use if you were looking for national news?

How about if you are searching for medical information?

A note on copyright and public domain images:

Images and other files and content on the Internet are protected in the same way as print materials and photographs. Use of digital images for purposes of alteration and display on the Internet has limited coverage under the conditions of fair use. [<http://www.templetons.com/brad/copymyths.html>] and [<http://www.copyrightwebsite.com/info/fairUse/fairUse.asp>].

Public Domain

[<http://www.copyrightwebsite.com/info/publicDomain/publicDomain.asp>]

items are those in which the copyright has been lost, has expired, or the author of the work makes no copyright claims to reproductions or enhancements of the work.

<http://www.unc.edu/~unclng/public-d.htm>

If you use an image of a person for reasons of making a profit, you are responsible for obtaining permission from the person or their heirs. If you use a trademark image, you must also get permission.

Copyright in websites:

[<http://www.copyrightwebsite.com/digital/webIssues/webIssues.asp>]

5. Using the Search Engine Math you read about, construct a search to find sites that contain images in the public domain. Use Google for this first search.
6. Do that same search across in AltaVista and Dogpile as well. Compare your top 10 hits. Do you get the same results?
 - How are they similar?

 - How are they different?
7. Try changing the search and see if you get different results. For example, if you did your first search as **+public +domain +images**, try a search with the phrase "**public domain images**" instead.

Do your results change?

8. Do a search for images related to Seattle on the web. Alta Vista has a way to just search for image media on the web. Can you locate other search engines with this same feature?
9. Find an image of the San Francisco skyline. Which search engine had the best image and what number was it in the results?
10. Now look for images you would like to use in a website of misinformation/missing information (Project 1 – will be released before this assignment is due) and save them for manipulation in Adobe

Photoshop later on. Remember to FTP all images to your Dante account so you'll have them for use later. How many images have you found and uploaded as of the time you print out your answers?

NOTE: Make sure that any image you select is in the Public Domain OR the copyright policy on the site where you find it states that you are allowed to use it for non-commercial purposes!!!!

Optional (no extra credit, just extra fun)

Is there any information about you on the web? If you just type in your full name, you'll probably get either no hits, or lots of hits that don't refer to you. Now that you know more about constructing search queries, construct the best query that you can which simultaneously

- of the existing pages which refer to you, finds the largest percentage of them (this is called the "recall")
- of the pages which it does find, has the largest percentage of them which refer to you (this is called the "precision").

Now answer:

- What search engine did you use?
- What was the exact query?
- What was the recall percentage, approximately?
- What was the precision percentage, approximately?