Announcement ...

https://www.youtube.com/watch?v=dU1xS07N-FA
Searching The WWW

Lawrence Snyder
University of Washington, Seattle
Google is not necessarily the first place to look!

- Go directly to a Web site -- www.irs.gov

Guessing a site’s URL is often very easy, making it a fast way to find information

- Go to your bookmarks -- dictionary.cambridge.org
- Go to the library -- www.lib.washington.edu
- Go to the place with the information you want -- www.npr.org

Ask, “What site provides this information?”
Google Advanced – Use It!

Find web pages that have...
- all these words:
- this exact wording or phrase:
- one or more of these words:

But don’t show pages that have...
- any of these unwanted words:

Need more tools?
- Reading level:
- Results per page:
- Language:
- File type:
- Search within a site or domain:

Advanced Search

Date, usage rights, numeric range, and more
Caution!

- In the next few slides, the general principles of keyword search are discussed ... Google and Bing “adjust” the results somewhat.
Boolean Queries

Search Engine words are independent

- Words don’t have to occur together
- 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
Searching strategies ...

- Limit by top level domains or format ... .edu
- Find terms most specific to topic ... ibuprofen
- Look elsewhere for candidate words, e.g. bio
- Use exact phrase only if universal, ... “Play it again”
- If too many hits, re-query ... let the computer work
- “Search within results” using “-” ... to get rid of junk
Queries, continued

- Once found, ask if site is best source
  - How authoritative is it?
  - Can you believe it?
  - How crucial is it that the information be true?
    - Cancer cure for Grandma
    - Hikes around Seattle
    - Party game
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 (stored in a database)
  - Query processor checks user requests against the index, reports on known pages [You use this!]

Only a fraction of the Web’s content is crawled

- We’ll see how these work momentarily
As you know, the Web uses http:// protocol. It’s asking for a Web page, which usually means a page expressed in hyper-text markup language, or HTML.

- *Hyper-text* refers to text containing links that allow you to leave the linear stream of text, see something else, and return to the place you left.

- *Markup language* is a notation to describe how a published document is supposed to look: fonts, text color, headings, images, etc. etc. etc.
Rule 0: Content is given directly; anything that is not content is given inside of tags

Rule 1: Tags made of < and > and used this way:

```
Attribute&Value
```

It produces: This is paragraph.

Rule 2: Tags must be paired or “self terminated”
Example

- Write HTML in text editor: notepad++ or TextWrangler
- The file extension is `.html`; show it in Firefox or your browser
Rule 3: An HTML file has this structure:

```html
<html>
  <head><title>Name of Page</title></head>
  <body>
    Actual HTML page description goes here
  </body>
</html>
```

Rule 4: Tags must be properly nested

Rule 5: White space is mostly ignored

Rule 6: Attributes (style="color:red") preceded by space, name not quoted, value quoted
Three Sides: Basics of HTML

- To put in an image (.gif, .jpg, .png), use 1 tag
  `<img src="skier.jpg" alt="Skier in Snow"/>
  
  Tag | Image Source | Alt Description | End

- To put in a link, use 2 tags
  `<a href="http://www.cs.uw.edu/cse120">Pilot</a>
  
  Hyper-text reference – the link | Anchor

- More on HTML (including good tutorials) at http://www.w3schools.com/html/default.asp
How to crawl the Web:

- Begin with some Web sites, entered “manually”
- Select page not yet crawled; look at its HTML
  - For each keyword, associate it with this page’s URL as in http://www.cs.uw.edu/cse120/example : downhill and http://www.cs.uw.edu/cse120/example : skiing
  - Harvest words from URL and inside <title> tags ...
  - For every link tag on the page, associate the URL with the words inside of the anchor text, that is, http://www.cs.uw.edu/cse120/ : pilot
- Save all links and add to list to be crawled
After crawling a page like

http://www.cs.washington.edu/education/courses/cse120/13wi/gallery.html

the crawler will associate many terms with the URL: Henry, Pong, Tron, ... as well as gallery, [from anchor] and cse120 [from URL]

Terms from URL and anchor are more important in describing the page
When the crawling is “done” (it’s never done), the result is an index, a special data structure that a query processor can use to look up your queries:

Henry: ..., www.cs.washington.edu/cse120/gallery.html, ...

Pong: ..., www.cs.washington.edu/cse120/gallery.html, ...

Tron: ..., www.cs.washington.edu/cse120/gallery.html, ...
Make A Query

- When Google gets the query

- It “ands” the two lists together, finding URLs that are on both lists
- It counts them up, records time, shows 10 hits
You want the most likely hits ... how does Google show you what you want?

- Page Rank – a mechanism to estimate the “importance” of a page; pages are listed by page rank, highest to lowest
Google has never revealed all details of the ranking algorithm, but we know ...

- URL’s are ranked higher for words that occur in the URL and in anchors
- URL’s get ranked higher if more pages point to them, it’s like: A links to B is a vote by A for B
- URL’s get ranked higher if the pages that point to them are ranked higher
Virtual Folders are a “crawling/querying” technology that helps you
- Mac: Smart Folders
- PC: Saved Folders
In both cases your files are “indexed”, that is, crawled, and the query you make results in a smart folder of the files that “hit”
It’s like Googling the stuff on your own computer
Query “thesis”

- The folder doesn’t exist ... it just contains links to the files shown

- Very convenient!
Search Engines ... A Summary

- A search engine has two parts
  - Crawler, to index the data
  - Query Processor, to answer queries based on index
- In the case of many hits, a query processor must rank the results; page rank does that by
  - “using data differentially” ... not all associations are equivalent; anchors and file names count more
  - “noting relationship of pages” ... a page is more important if important pages link to it

Google, Bing, Yahoo and other Search Engines Use All of These Ideas