

Web Pages and Search Engines

Key Learning Goals

After the web unit, you will be able to:

- connect your knowledge of file systems and networks to web pages and URLs
- explain how a web search engine finds and indexes new pages

reminder: URLs

- <http://www.google.ca/index.html>
 - **http** refers to a protocol for transferring files
 - **www.google.ca** is a domain name
 - **index.html** is a file name (at the google domain)
- <http://www.ugrad.cs.ubc.ca/~cs101/current-term/Labs/Getting-Started/index.html>
 - `~cs101/current-term/Labs/Getting-Started/index.html` specifies the main web page for the lab. It's the file named "index.html" in:
 - the "cs101" directory's
 - "current-term" subdirectory's
 - "Labs" subdirectory's
 - "Getting-Started" subdirectory

web page links

- a link has:
 - an *anchor*: the underlined text you click on
 - a *hyperlink reference*: the URL of the web page you see when you click on the link

So, if we think of each page as a "node" and each link as an "edge" connecting to another "node" ...

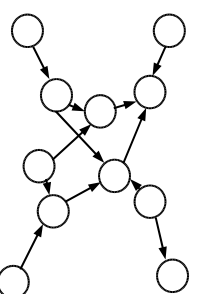
the world wide web: web pages and links

- web pages are files stored on servers (computers)
- each web page has a URL made of "http" (meaning it's a web page), the server's name, and the file's location on the server
- many web pages are written in a language called *HTML (HyperText Markup Language)*, which we'll learn soon
- HTML pages can link to other web pages, indicating those other pages' URLs!

"networks" of web pages

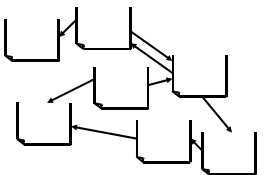
If we think of each web page as a "node" and each link as an "edge" between nodes, a group of web pages starts to look a lot like our diagram of the Internet...

In CS, we can think of them both as structures called "graphs": networks of nodes and edges.



organization of the web

- the web is a network, with **directed links**
 - nodes are web pages or documents
 - nodes and links are constantly changing
- *search engines* like Google find information on the web by taking advantage of that networked structure



search engine

- a search engine is a collection of computer programs for finding information on the WWW
- a typical search engine has three components:
 - a *crawler* (spider, or robot)
 - a *query processor*
 - an *interface* (typically a web page)

crawler, query processor, interface

- The *crawler* creates a (*keyword, URL*) table; keywords are taken from the title of the document and from the *anchors* of links to the document.
- The *query processor* uses this table to find URLs that match the keywords entered by the user in the search engine *interface*.

crawler example

www.attractions.ubc.ca

Attractions and Recreation at UBC

Chan Centre
...
...

table	
attractions,	www.attractions.ubc.ca
recreation,	www.attractions.ubc.ca
ubc,	www.attractions.ubc.ca
chan,	www.chancentre.com
centre,	www.chancentre.com
performing,	www.chancentre.com
arts,	www.chancentre.com
about,	www.chancentre.com/about_f.html
chan,	www.chancentre.com/about_f.html

www.chancentre.com

Chan Centre for the Performing Arts

About the Chan
...
...

crawler example

www.attractions.ubc.ca

Attractions and Recreation at UBC

Chan Centre
...
...

table	
attractions,	www.attractions.ubc.ca
recreation,	www.attractions.ubc.ca
ubc,	www.attractions.ubc.ca
chan,	www.chancentre.com
centre,	www.chancentre.com
performing,	www.chancentre.com
arts,	www.chancentre.com

work of crawler never ends...

- ... since pages are constantly being modified, deleted, or added to the web!
- Next week, you'll create a page!
- I'll link to your page from the class home page.
- Try out Google occasionally to see if and when your page is returned when you do a web search.