The World Wide Web

Lecture 7 - COMPSCI111/111G

"On the Internet, nobody knows you’re a dog."
Today’s lecture

- Understand how the WWW works
- Understand how search engines work
- The implications of search engines
The WWW project

- Tim Berners-Lee built the first website in 1991 at CERN.
The WWW project

- In June 1993, Mosaic was released; the first widely used web browser

- By Oct 1993, there were 500 web servers around the world
  - By this point, Berners-Lee realised the WWW had to be freely available so he convinced CERN to make the source code public

- In 1994, Berners-Lee established the World Wide Web Consortium (W3C), which created standards for the WWW.
Evolution of the Web

- 1994: Netscape Communications and Yahoo! founded
- 1995: first version of Microsoft Internet Explorer released
- 1994: Amazon founded by Jeff Bezos
- 1998: Google founded by Brin and Page
- 1997-2001: “Dot-com” boom and bust
- 2004: shift to ‘Web 2.0’ (eg. wikis)
Some terms

- **Webpage**: a document on the WWW that is usually accessed through a web browser
- **Website**: a collection of webpages usually on the same topic or theme
- **Web browser**: application software used to access content on the WWW
- **Web server**: a computer with software that makes files available on the WWW
Uniform Resource Locator (URL)


- Protocol: https
  - Other common protocols: ftp, http

- Domain: www.cs.auckland.ac.nz
  - Can be a domain name or an IP address

- Path on server: /~andrew/

- Resource: teaching.html
HTTP

- HyperText Transfer Protocol; used by web browsers to request resources (eg. webpages, images, sounds) from a web server

- There’s also HTTPS = HyperText Transfer Protocol Secure
  - Encrypts the HTTP connection using TLS (Transport Layer Security)
  - Becoming essential for websites to use HTTPS to keep user information secure
Find IP address of www.google.com

CLIENT

GET /index.html HTTP/1.1

HTTP/1.1 200 OK

SERVER

GET /img/logo.jpg HTTP/1.1

HTTP/1.1 404 NOT FOUND
Other parts of the WWW

- **Proxy**: sits between client and server so it can intercept and process requests

- **Cache**: stores recently requested resources so they can be accessed quickly
  - A proxy can use a cache to store recent requests, enabling it to process requests faster

- **Firewall**: prevents unauthorised access to a private network
Search engines

- A website that helps a user to search for information on the WWW

- Software indexes content on the web. This index is used to build a list of results based on the search terms entered by the users
  - **Indexing**: organising data so that it is easier to search

- Popular search engines include:
  - Google
  - Bing
  - Yahoo search
  - DuckDuckGo
How do search engines work?

- Spiders crawl across the WWW to scan webpages
  - Spiders are programs that follow links and gather information from webpages

- The search engine’s index is updated with information gathered by the spiders
Filter bubble

- Occurs when a search algorithm offers personalised results, which limits the diversity of information presented to the user
  - Examples include Facebook’s News Feed and Google’s personalised search results

- Personalised search results can help people to find relevant information

- However, it also risks isolating people within their own bubble of information
Questions

- What problem did Tim Berners-Lee want to solve using the Web?
- What is the difference between a firewall and proxy?
Answers

- What problem did Tim Berners-Lee think he could solve using the Web?
  - Sharing information between researchers at CERN

- What is the difference between a firewall and proxy?
  - Firewall: prevents unauthorised access to a network
  - Proxy: intercepts and processes requests from clients and servers
Summary

- The WWW was designed to be a system to share information
  - It has become a system for creating and sharing a variety of content
  - Key protocol on the WWW is HTTP

- Search engines use an index of the WWW to provide results based on search terms