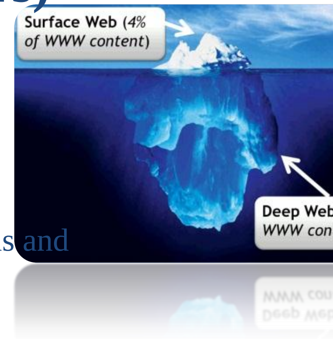
C.2 Searching the web (6 hours)



C.2.1 Define the term search engine.

Search Engines are programs which search documents for specific keywords and return a list of results.

Typically, Web search engines work by sending out a spider/webcrawler (automated program which ‘crawls’over the web to fetch as many documents as possible. Another program, called an *indexer*, then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query.

Web browsers such as Google Chrome, Firefox and Internet Explorer use Web search engines such as Google, Bing and Yahoo.

http://www.webopedia.com/DidYouKnow/Internet/HowWebSearchEnginesWork.asp

C.2.2 Distinguish between the surface web and the deep web.

The surface web or visible web is searchable and accessible by the public using standard search engines.

The deep web or invisible web is often accessible only via individual search pages and often many layers deep with complex URLs.

Often password protected or requires a subscription. The deep web is much larger than the surface web.

Lexibot is an example of a deep web specific search engine.

http://www.answers.com/topic/surface-web-1

http://www.answers.com/topic/deep-web

C.2.3 Outline the principles of searching algorithms used by search engines.

PageRank is a ranking system used by Google which indicates the importance/popularity of a page on the web.

It uses a numerical system of ranking and takes into account the number of links to a page to indicate importance, the PageRank value is one factor which decides

Which position a web page is displayed in search results.

http://www.webworkshop.net/pagerank.html

http://www.whitehatworks.com/google_latest_algorithms.asp

HITS

Hyperlink Induced Topic Search - a precursor to PageRank

http://en.wikipedia.org/wiki/HITS_algorithm

C.2.4 Describe how a web crawler functions.

Spiders/bots gather the information for sorting/indexing. This is known as Web crawling.

Spiders begin with lists of popular pages and then follow any links on those pages. Multiple spiders can be used simultaneously (Google started with four) and can ‘crawl’ over thousands of pages per second.

http://computer.howstuffworks.com/internet/basics/search-engine1.htm

C.2.5 Discuss the relationship between data in a meta-tag and how it is accessed by a web crawler.

Meta tags allow the owner of a page to specify key words and concepts under which the page will be indexed. This can provide a ‘guide’ for a search engine.

Normally created by the page owner. Point to note is that meta tags are normally created by the page owner and may be incorrect or deliberately designed to push the page up the results list by using a popular topic in the meta tag which may have little to do with the actual contents of the page.

C.2.6 Discuss the use of parallel web crawling.

The use of multiple spiders simultaneously.

C.2.7 Outline the purpose of web-indexing in search engines.

Indexing - Most search engines store more than just the word and URL. An engine might store the number of times that the word appears on a page. The engine might assign a weight to each entry, with increasing values assigned to words as they appear near the top of the document, in sub-headings, in links, in the meta tags or in the title of the page. Each commercial search engine has a different formula for assigning weight to the words in its index. This is one of the reasons that a search for the same word on different search engines will produce different lists, with the pages presented in different orders.

This information is usually encoded to save storage space.

C.2.8 Suggest how web developers can create pages that appear more prominently in search engine results.

Meta tags, links.   SERPS (Search engine return pages)

Link farms

Research Opportunity. Investigating search engine algorithms may help, what information is indexed by webcrawlers…

http://computer.howstuffworks.com/search-engine-optimization.htm

C.2.9 Describe the different metrics used by search engines.

Metrics indicate time taken, number and quality of returns. This could lead to the possibility of manipulation or 'queue jumping'.

http://computer.howstuffworks.com/search-engine-optimization.htm

https://support.google.com/analytics/answer/1032321?hl=en-GB

C.2.10 Explain why the effectiveness of a search engine is determined by the assumptions made when developing it.

Students will be expected to understand that the ability of the search engine to produce the required results is based primarily on the assumptions used when developing the algorithms that underpin it.

LINK Connecting computational thinking and program design.

C.2.11 Discuss the use of white hat and black hat search engine optimization.

White hat http://computer.howstuffworks.com/search-engine-optimization2.htm the goody web page designer….

Black hat http://computer.howstuffworks.com/search-engine-optimization3.htm the baddy web page designer….

Webmasters can try to trick search engines into listing their Web pages high in SERPs. Various dodgy tricks on the link above.

AIM 8 Developers of search engines should have a moral responsibility to produce an objective page ranking.

C.2.12 Outline future challenges to search engines as the web continues to grow.

issues such as error management, lack of quality assurance of information uploaded.
AIM 9 Develop an appreciation that search engines will need to evolve to remain effective as the web grows.

Research Opportunity.