IMS1603 Lecture 10

Searching the web

Today’s Lecture

• Some thoughts on searching the web
• A look at the tute/lab exercises
• Thinking a bit more about Assignment 1
• Another unit objective

Searching the web

The earliest search tools for the World Wide Web evolved out of individuals’ bookmark listings
An early success

- The most famous of the early search tools was Yahoo!
- It was first known as ‘Jerry and David’s guide to the World Wide Web’
- Yahoo’s Directory was manually compiled

Search engines proper ...

began with automated programs (‘spiders’ or ‘bots’) that hunted the WWW for information by following hyperlinks, and stored their findings in searchable databases

Hyperlinks

- the thread of interconnections between web pages - are what make this process of web searching possible
From August 1994

"The Carnegie Mellon University Center for Machine Translation announces the availability of the Lycos WWW search engine
http://fuzine.mt.cs.cmu.edu/mlm/lycos-home.html

Lycos provides probabilistic retrieval of over 300,000 WWW documents, with more being added every week. In addition to document pointers, Lycos provides match score, WWW links, an outline, keyword list and excerpt for each of the top 50 documents matching your query.

--Dr. Michael L. Mauldin Carnegie Mellon University Center for Machine Translation Pittsburgh, PA 15213 fuzzy@cmu.edu
http://fuzine.mt.cs.cmu.edu/mlm/home.html
http://groups.google.com/groups?q=fuzzy@cmu.edu&output=gplain

Search engine attributes

• Many of the early engines boasted powerful new features for the Internet, such as:
  – Full text searching
  – A broad coverage of web sites

Searching the web

• The shakeout of the so-called ‘New Economy’ in the late 1990s saw some search engines vanish, others mutate into something else

• After the dust had settled, one major contender would dominate the search engine world
The current popular champion

- Google combines:
  - full text searching
  - one of the broadest coverages of the WWW
  - relevancy ranking

Google on relevancy ranking

PageRank Explained

'PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves “important” weigh more heavily and help to make other pages “important.”

'Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Of course, important pages are not automatically ranked higher. Instead, relevancy-ranking algorithms determine the relative importance of different Web pages and provide a means by which Google can determine your preferences in a web search. Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query.'

http://www.google.com/technology/index.html

Google on relevancy ranking

'Google's complex, automated search methods preclude human interference. Unlike other search engines, Google is structured so no one can purchase a higher listing or commercially alter results. A Google search is an honest and objective way to find high-quality websites, easily.'

One critic (and rival) of Google

‘Google’s key failing is in its old-fashioned search technology, based on statistical algorithms that don’t actually understand the context of a search. Every word in the English language has, on average, two-and-a-half related senses — consequently, search results are returned that can be unrelated to each other apart from word association.’

(Ian Saunders, quoted in Winder 2005: 119).

Searching the web

• Which raises some interesting questions about searches on Google:
  – Google bombs
  – Paid placement (but not paid inclusion)

Searching the web

An interesting experiment: natural language queries

• eg Ask Jeeves
Meanwhile …

What you see is not necessarily what you get

http://www.ihelpyou.com/search-engine-chart.htm

Searching the web

- While their use is widespread, search engines continue to present problems for many users
  - Weaknesses of keyword searching
  - Recall and precision
  - The authority of information retrieved
  - The breadth of coverage

What is the ‘invisible web’?

It’s made up of those millions and millions of web pages which are not captured by existing search engines
Exercises for Lab 4

• Comparing search engines
  – Carry out some searches using a variety of platforms, and evaluate the results
  Let’s try one or two now.

Exercises for Tute 4

Analyse some documents:
  – Form
  – Format
  – Medium
  – Context
  – Authority
  – Content
  – Purpose
  – Technology(ies)
  – Accessibility

Thinking about Assignment 1

• Part 1: evaluate the document provided
  – What apparent statements of ‘fact’ does it contain?
  – How might you verify them?
  – What kinds of records would you turn to?
Biography of Alan Alexander Milne (1882 - 1956)

'Scots by birth, Alan Milne spent his childhood in London, where his father was a preparatory schoolmaster. His early education owed much to the skills of a young teacher and mentor — H.G. Wells — years later, Milne described Wells as "a great writer and a great friend." He continued his education at Westminster School and Trinity College, Cambridge. He bequeathed his original handwritten manuscripts of Winnie-the-Pooh and The House at Pooh Corner to the College Library. While an undergraduate at Cambridge he edited Granta for a year — his first literary efforts were published in the humorous magazine Punch, where a month after his twenty-fourth birthday he started work as Assistant Editor, remaining there until the outbreak of the First World War.'

http://www.pooh-corner.com/biomilne.html

Unit objectives

- At the completion of this unit the students will:
  - have knowledge of:
  - have an understanding of:
  - have the skills to:
  - have developed attitudes which enable them to:

- have the skills to:
  - evaluate information and its sources critically
  - identify particular information needs
  - evaluate ICTs and IM tools in terms of meeting user needs
  - use basic ICT and IM tools to create, represent, store, access, retrieve and use information
have the skills to:

- use basic ICT and IM tools to create, represent, store, access, retrieve and use information

How about the Voyager catalogue as one example?

Further reading

UNSW Press.
A. Hiriart (1994) When we were very young. London: Methuen.
Hill/Osborne.
http://www.searchengineshowdown.com/
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html

Next lecture

Documentary form and the impact of ICT