Computer Science

COMPSCI 111 SC - Lecture 27

October 2003

As We May Think

Hypertext and the World Wide Web

Introduction

The WWW is a fairly recent phenomenon, yet the underlying structure has had a long history of development which is still underway today. The current structure of the WWW has serious flaws which reduce its effectiveness as a hypermedia system. The interface through which we access the WWW is fluid, always shifting under competitive pressure. Through all the changes one thing remains consistent - the public interest.

Hypertext

Vannevar Bush (Science Advisor to Roosevelt during WWII) proposed the Memex system in his 1945 article "As We May Think". This system was a conceptual machine which could create information trails. These trails were links between related texts and illustrations which could be used for reference. Bush felt that such a machine would greatly increase learning, memory and knowledge.

Inspired by Bush, Douglas Englebart (working in 1963 at SRI), proposed a system which cross-referenced related documents across a network (later he invented the idea of a mouse and screen pointer for GUIs). Ted Nelson's work in 1960, was far more ambitious. His project, Xanadu, was designed as a document universe, where everything ever written would be stored and referenced with crosslinks to related information. He coined the term "Hypertext" in 1965. After much funding, the Xanadu project is still underway today, still under the leadership of Ted Nelson. His continuing advocacy of a 30 year old project has resulted in some criticism of Nelson

> *"Xanadu, the grandest encyclopedic project of our era seemed not only a failure but an actual symptom of madness." - Wired magazine*

Tim Berners-Lee began work on the WWW project at CERN in 1989. The European Particle Physics Laboratory had already helped shape the Internet by supporting the TCP/IP protocol. Tim Berners-Lee visualised a hypertext system which would encourage collaborative research. Contributors would have to have world-wide access through the Internet, and would be able to easily add to the database of knowledge and cross-reference other documents.

Development

The WWW was fully operational at CERN in 1991. At this time, only text was used, and the only browser used a CLI. The WWW looked a lot like other aspects of the Internet (Usenet, E-mail etc;). By 1992, the National Centre for Supercomputer Applications (NCSA) released the Mosaic browser with a GUI interface. A year later (in 1993), GUI versions of the browser were released for home computers, both PC and Macintosh versions. Since that time, the ease of use has allowed the general public to become involved.

The Underlying Structure

The protocol used by the WWW is the Hypertext Transfer Protocol (HTTP). The hosts which support HTTP are able to talk to each other, and pass documents back and forth, forming the basis of the WWW. Note that the WWW is not the same as the Internet, but rather a collection of servers (a subset of Internet hosts), which support HTTP.

Client-Server Model

The WWW uses a client-server model as the basis of communication. In this model, the client (browser) runs on the local (or user's) machine, and the server runs on one of the host machines. The client is responsible for displaying the information in the documents, displaying and maintaining hypertext links in a intuitive manner and negotiating formats of information with the server. The server must negotiate formats with the client, send information in the requested format and manage the nodes of information on the host machine upon which it is running.

Structural Problems

The hypertext structure developed for the WWW has some major flaws. The most obvious of these is the problem of "dangling links". These occur when a user moves or deletes a document available on the WWW. All the other documents which contain hypertext references to the first document are now left with links which refer to a page that no longer exists.

The lack of facilities for accounting inherent in the structure (due to its anonymous nature) make it undesirable for the publisher to publish professional quality work on the WWW. Publishers are generally uninterested if there is no way to make a profit. This has further encouraged people to look to advertising in order to maintain the web sites.

Due to the decentralised structure of the WWW, finding relevant information can be difficult. There is no central database or index of information, and no way to categorise information based on quality. Search engines are fully automated, and so they tend to index information poorly.

Search Engines

Many attempts have been made to index the WWW. This is usually achieved through the use of an automated program which recursively accesses all the pages on the WWW. For each page, the program will attempt to extract out the most important information from the document and send it back to a central database. This database can be searched for key words, and will display a list of pages which contain that word. The searchable database is known as a search engine.

WWW Demographics

The WWW is still predominantly used by males (69%) rather than females (31%). The average age is 35 years old, and has been steadily increasing over the past 5 years. Approximately half are married (46%) and one third single (37%). One in five users (20%) use the WWW for more than 20 hours per week, and almost a third spend 10-20 hours a week browsing. The most common use of the WWW is simply browsing (77%), followed by entertainment (64%), education (53%), work (51%) and shopping (19%). In recent surveys of users, the most common concern expressed by users is the issue of censorship (36%), followed by privacy issues (26%) and difficulty in navigation (14%). A telling statistic is that 37% claim to use the WWW instead of watching TV on a daily basis. Equally interesting is the response showing that users tend to spend as much time using e-mail as they do using the phone. *(Statistics from GUV's 6th WWW User Survey)*

Navigating through Cyberspace

Imagine a library where the books are placed on the shelves in no particular order. The only way to access them is through an index containing all the words which appear inside all the books. Imagine that this library also contains all the junk mail produced, and all the business documents produced by companies. What you have imagined is the World-Wide Web as it stands today. Navigating through the mass of information on the WWW can be a demoralising and unsatisfactory experience. The information which is available is usually difficult to find, even if you know exactly what you are looking for. The most frustrating aspect of the WWW and the thing which prevents most people from making effective use of the WWW is the lack of direction. Is far too easy to get lost in cyberspace, confused about where you have come from and where to go to next.

In order to master this new electronic environment you must learn how to use the tools for finding information on the WWW. The most important aspect is to choose the right tool for the job. There are many specific indexes, or databases which help to index specialised subjects (such as the Internet Movie Database). If you find any databases related to your own interests, then bookmark them, and build up a set of resources tailored to your tastes. If you are looking for general information about a topic, but don't have any specific queries (eg; you are generally interested in watersports), then using one of the subject catalogs will help to guide you to appropriate pages. Only use search engines when you are searching for specific information.

Search Engines or Subject Catalogs?

A subject catalog (sometimes called a directory or guide) is a manually-created catalog of sites on the WWW. This means that a person has created categories and then a team of people review a page, and if they consider it to be a worthwhile resource, they will include a link to it in the appropriate place in the categorical index. This means that all the sites within a subject catalog have been screened for information content by a human, so you are unlikely to find personal home pages and the like. If the editors of the catalog do not like the content of a page, then they will not add it to the directory. These

directories usually only index a small portion of the WWW, but the pages which are listed are usually of high quality.

A search engine uses an automated program to index all the pages in the WWW in a single enormous database. So how does a computer program know what the content of a page is? Each search engine uses a different approach to this problem. Most of them examine the page, and record keywords from the page. They are likely to take account of the number of times a word appears, and it's position in the page. The title and headings of a page are often highly regarded as indicative of the content of the page by an indexing program.

Using the URL for searching

The "branding" of a web site is possibly the most important aspect. Each web site has a unique address. If this address is memorable, then you can find the site easily and reliably. Sites like "Yahoo!" (http://www.yahoo.com) or "Amazon" (http://www.amazon.com) have succeeded in this contest for a place in the consciousness of the casual Internet user.

Understanding how web sites are named gives you a big advantage in searching for sites. Searching for a site by trying to guess the URL is a remarkably quick and easy way to get started. Try a few guesses to start with and see if any are useful. You can always resort to a search engine. Example: If you are looking for tourist information about NZ where would you look? Try: *www.tourist.co.nz, www.tourism.govt.nz*, *www.touring.org.nz*. If none of these are successful, then perhaps the NZ *govt* site would have links to other tourist sites. Try *www.govt.nz*. Maybe each local govt body would have information about their area. If you find a site, but its not quite what you want, then look for Links to other related sites.

Summary:

1. Use specific databases if they exist (and if you can find them)
2. Use catalogs to find general subject information
3. Use search engines for specific queries

Simple Searching using Alta Vista

Always try a simple search using natural language. Type a word or phrase or a question (for example, "weather Boston" or "what is the weather in Boston?"), then click Search (or press the Enter key). If the information you want from this sort of query isn't on the first couple of pages, try adding a few more specific words (like "June", or "today", or "weather report")

Required and rejected terms

Often you will know a word that will be guaranteed to appear in a document for which you are searching. If this is the case, require that the word appear in all of the results by attaching a "+" to the beginning of the word (for example, to find an article on pet care, you might try the query dog cat pet +care). You may also find that when you search on a vague topic, you get a very broad set of results. You can quickly reject results by adding a term that appears often in unwanted articles with a "-" before it (for example, to find a recipe for oatmeal raisin cookies without nuts try: oatmeal raisin cookie +recipe -nut* -walnut*).

Phrases

If you know that a certain phrase will appear on the page you are looking for, put the phrase in quotes. (for example, try entering song lyrics such as "you ain't nothing but a hound dog"). Using quote marks ensures that the words appear in the exact same order that you have specified, otherwise you will find *all* documents containing *any* of the words: dog, hound, nothing, ain't, you, a, but

Case Sensitivity

Use only lower case unless you want your search to be case sensitive. If you search for Coffee, you'll get only documents that include that word with just that capitalisation. If you search for coffee, you'll get any page with that word.

Wildcards

Use an asterisk (*) to broaden your search. To find any words that start with gold, use gold* to find matches for gold, goldfinch, goldfinger, and golden. Use this if the word you are searching for could have different endings (for example, don't search for dog, search for dog* if it could be plural).

Special Functions for web searches using Alta Vista

AltaVista doesn't just search text. Here are all of the other ways you can search on the net:

Hypertext Links

You can find pages which contain a word or phrase within the test of a hyperlink by using anchor:text. For example, anchor:"Click here to visit AltaVista" would find pages with "Click here to visit AltaVista" as a link.

Destination URLs

You can find all pages which link to a destination URL. This may be useful if you wished to find all pages which have links to a page you are interested in (such as your own home page). Use link:URLtext to find all such pages. For example, link:altavista.digital.com finds all pages which link to the Alta Vista search engine.

Images

You can search for images by using image:text. For example, image:elvis looks for pages with images called elvis. Note that the search is based on the name of the image file, so use short names without spaces (since filenames are likely to be single words less than 8 characters long).

Titles

If you know the title of the page you are looking for (ie; the name which usually appears in the title bar of the browser), then you could use title:text. For example, a search for title:Elvis would find pages with Elvis in the title.

Advanced Features of Alta Vista

When a general search has not been successful, and you are looking for specific information, then an advanced query may provide better results. Advanced search is for very specific queries and not for general searching. Almost everything you need to do can be done more quickly and with better results through the simple form. Remember, when you use the advanced search form, you control the ranking and if the ranking field is left blank, no ranking will be applied and the results will be in no particular order.

Boolean Operations:

Note that the + and - operators do not work in an advanced search. You should use the Boolean keywords AND, OR, NOT, and the operator NEAR. Each of these operators has a shortened form, respectively &, |, !, and ~.

AND, &

Finds only documents containing all of the specified words or phrases. *Mary AND lamb* finds documents with both the word *Mary* and the word *lamb*.

OR, |

Finds documents containing at least one of the specified words or phrases. *Mary OR lamb* finds documents containing either *Mary* or *lamb*. The found documents could contain both, but do not have to.

NOT, !

Excludes documents containing the specified word or phrase. *Mary AND NOT lamb* finds documents with *Mary* but not containing *lamb*. NOT cannot stand alone--use it with another operator, like AND. For example, AltaVista does not accept *Mary NOT lamb*; instead, specify *Mary AND NOT lamb*.

NEAR, ~

Finds documents containing both specified words or phrases within 10 words of each other. *Mary NEAR lamb* would find the nursery rhyme, but likely not religious or Christmas-related documents.

Ranking results:

To rank matches, enter terms in the Ranking field; otherwise, the results will appear in no particular order. You could enter words that are part of your query or enter new words as an additional way to refine your search. For example, you could further narrow a search for *COBOL AND programming* by entering *advanced* and *experienced* in the ranking field

Using Search Engines

In order to find information quickly, you need to practice using a search engine until you are confident with it. Spend a little time using a few different search engines, then pick your favourite one and learn how to use the advanced features. Each search engine indexes pages in a slightly different way, and use can use different techniques to narrow your search with each of them. It is worthwhile finding out how the search engine you use actually indexes pages, since that will give you insight into the results of your searches (ie; why certain pages are near the top of the list). The search engines which are most highly recommended are Google, Hot Bot, and Alta Vista, but each person is advised to try a range and select the one which you are most comfortable using.

Resources on the WWW

There are many resources available on the WWW. Here are a variety of commonly accessed pages which provide a good introduction to the type of resource material which can be found.

Common Search Engines

Google	http://www.google.com
Alta Vista	http://www.altavista.com/
Hot Bot	http://www.hotbot.com/
Lycos	http://www.lycos.com/
Excite	http://www.excite.com/

Common Directories (Guides)

Galaxy	http://galaxy.einet.net/
Yahoo!	http://www.yahoo.com/
LookSmart	http://www.looksmart.com/

Library Resources

Electric Library	http://www.elibrary.com/
Online Books	http://digital.library.upenn.edu/books/

Shopping

Amazon Books	http://www.amazon.com/
Blackstar	http://www.blackstar.co.uk/

Reference

Webopedia	http://www.webopedia.com
NetLingo	http://www.netlingo.com/

Entertainment & News

Internet Movie Database	http://www.imdb.com/
New York Times	http://www.nytimes.com/
Satire Wire	http://www.satirewire.com

Health

Kids Health	http://kidshealth.org/index2.html

Software

Shareware	http://www.shareware.com/

New Zealand

AUSA	http://www.ausa.auckland.ac.nz
New Zealand Herald	http://www.nzherald.co.nz/
Trade and Exchange (NZ)	http://www.te.co.nz/
Auckland Public Library	http://www.akcity.govt.nz/library/

Author: Andrew Luxton