

Artificial Intelligence in Internet Search Engines.

E2025 Assignment

November / December 1999

Over the last few years, use of the Internet has grown exponentially, to the point where now it is hard to look anywhere and not see an Internet address or URL. The main use of the Internet is as a source of information, but as the Internet (in general) has no structure or control, we must search for any information that we might wish to seek.

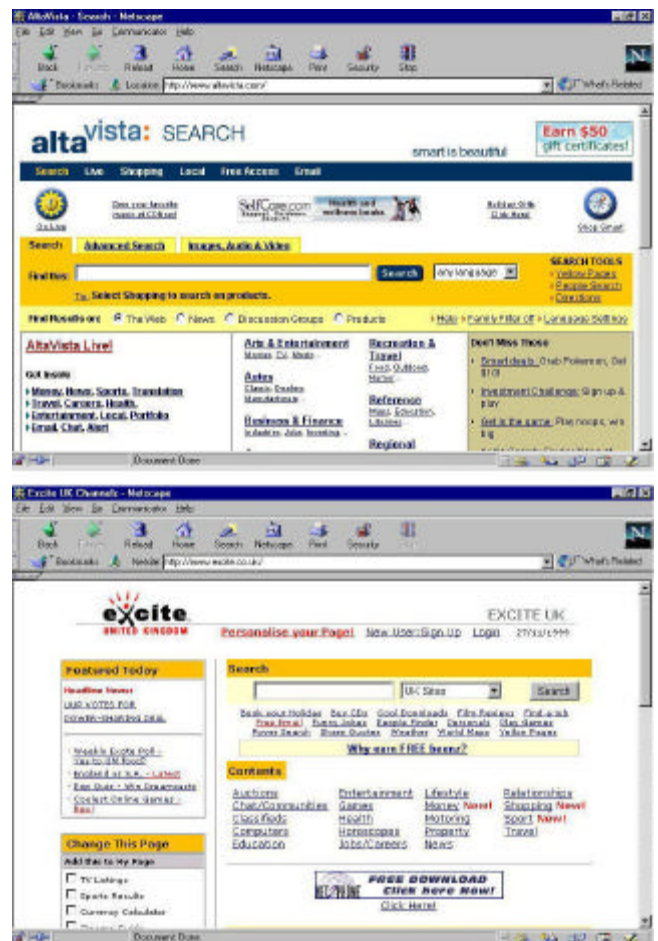
To do this, we have two options:

- (1) Use a subject directory like Yahoo, which relies on humans to submit web pages to its central database or index;
- (2) Use a search engine, which discovers web sites by itself.

How these search engines collect information about the millions of web sites that exist is the subject of this assignment.

When a search engine such as AltaVista or Excite returns matches for a search, how does it know that the sites it presents exist? The answer lies in the “spiders”, “robots” or “crawlers” that the search engine uses to crawl through the Internet, looking for sites to add to the search engine’s index.

Given a web page to start from, the spider searches for links on the web page, which may be internal or external links. For each page that it encounters, following a particular link, the spider records some information about the page, and stores this information in the search engine’s index. The information it stores will typically be the URL of the page and a short description of what the page’s purpose is, perhaps using meta tags, the first few lines of the page, and any previous information it has on the page. Note that the author of a web page may exclude the spider from indexing the page by placing a special file in the same directory as the HTML file.



The Internet is constantly expanding, and there is never enough time or computing power to index all Internet pages comprehensively. Indeed, although the best search engines claim to index 90% of all Internet Web Sites, this only accounts for about a third of all web *pages* at any given time.

Because we cannot index *all* the material that exists on the Web, we must reach a compromise of only indexing some of it. Using the depth first method, we would cover some subjects comprehensively, while the breadth first method would cover many subjects in limited detail. If we were only allowed to index four pages in the example on the previous page, then the depth first method would give us detailed information about Liverpool, while the breadth first method would give us some general information about all the semi-final teams.

Obviously, the optimum solution is to cover as many subjects in as much detail as is possible. That is why search engine spiders use a combination of both depth and breadth first indexing. Perhaps a *beam-first* or *best-first* method is used, where the scoring function is something like how frequently the page is updated or how many pages exist on the web site.

Spidering the web is a time consuming process. It is not a matter of indexing every page once, as many pages are regularly updated. Returning to a page after some time interval, the spider would more than likely encounter different material and different links to pursue on its journey around the web. When designing a spider, we must take into account that the Internet changes and expands on a daily basis. Unfortunately, executing a simple search on today's date and a date last month, and then comparing the amount of matches obtained, will show you that search engines do not return current information as they should.

We must also be wary of situations where we have a lot of links to the same page. An example of this would be a site that has a link to its home page on the top of every page on the site. Blindly following links, we could end up indexing the same page perhaps hundreds of times. While a human would instantly recognise that we have already visited a page, we must make sure our spider, essentially a piece of computer software, can ignore a page that we have already visited. This saves a lot of time and prevents looping.

In summary, to deal with the mountain of information on the Internet, we will always need artificial intelligence techniques, as a human could never deal with the volume of information available. In the ever changing landscape of the World Wide Web, we need the latest, fastest techniques to make searching and browsing the Internet as easy and as accessible as possible. Over the coming years, development in search engine AI could mean integrated voice recognition; fast, comprehensive indexing, and much better filtering of results.

Bibliography

HENRY, M; HUGHES, J. (1999), Effective Internet Use (online)
Available from: <http://www.libraries.psu.edu/crsweb/lifesci/medinfo/howsrch.html>
(Accessed 1st December 1999)

KOSTER, M. The Web Robots FAQ... (online)
The Web Robots Pages
Available from: <http://info.webcrawler.com/mak/projects/robots/faq.html>
(Accessed 1st December 1999)

PROSISE, J (July 1996) Crawling the Web (online)
PC Magazine
Available from: <http://www.zdnet.com/>
(Accessed 1st December 1999)

SULLIVAN, D (1999) Beyond The Hype: Dissecting AltaVista's Claims (online)
Search Engine Watch
Available from: <http://www.searchenginewatch.com/sereport/99/11-avclaims.html>
(Accessed 1st December 1999)

SULLIVAN, D. (1996-1998) How Search Engines Work (online).
Search Engine Watch.
Available from: <http://www.searchenginewatch.com/webmasters/work.html>
(Accessed 27th November 1999)

WINSHIP, I.R. (1995) World Wide Web searching tools - an evaluation. (online)
Information Services Department, University of Northumbria at Newcastle, UK
Available from: <http://bubl.ac.uk/journals/lis/oz/vine/n09995/winship.htm>
(Accessed 27th November 1999)