Upgrading search strategy skills: unit three: developing search strategies for searching some of the indexes of the world wide web: a continuing professional development for library and information service managers

Gulten S. Wagner

*Edith Cowan University*

Follow this and additional works at: [https://ro.ecu.edu.au/ecuworks](https://ro.ecu.edu.au/ecuworks)

Part of the *Library and Information Science Commons*


This Book is posted at Research Online.
Edith Cowan University

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study.

The University does not authorize you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following:

- Copyright owners are entitled to take legal action against persons who infringe their copyright.
- A reproduction of material that is protected by copyright may be a copyright infringement.
- A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
Upgrading Search Strategy Skills

Unit Three

Developing Search Strategies for Searching Some of the Indexes of the World Wide Web

A Continuing Professional Development Course for Library and Information Service Managers

Dr Gülten S. Wagner
Edith Cowan University
Copyright © 1996

Edith Cowan University

All rights reserved. Except for the quotation of short passages for the purpose of criticism and review, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission of Edith Cowan University.

ISBN O-7298-0283-3
# TABLE OF CONTENTS

## INTRODUCTION

- Unit Outcomes ................................................................. 4
- Library Competency Standards ........................................ 5
- Legal and Ethical Dimensions of Search Strategy Development Over the World Wide Web ......................................................... 5

## 01 WHAT IS A SEARCH STRATEGY?

- Objectives ........................................................................ 6
- Competency Standards ...................................................... 6
- Introduction ...................................................................... 6
- The Search Process ............................................................ 7
- Translating the Information Request - Identification of Key Concepts .......................................................... 7
- Boolean Logical Connectors .............................................. 9
- Stopwords ........................................................................ 11
- Precision and Recall ........................................................ 12
- Strategies for Narrowing and Broadening a Search ............ 12
- Truncation ....................................................................... 13
- Proximity or Contextual Operators .................................... 14
- Summary ......................................................................... 17

## 02 WHAT IS A WORLD WIDE WEB SEARCH?

- Objectives ........................................................................ 18
- Competency Standards ...................................................... 18
- Introduction ...................................................................... 19
- Veronica (Very Easy Rodent-Oriented Net-Wide Index to Computerised Archives) and Jughead (Jonzy's Universal Gopher Hierarchy Excavation and Display) - Two Internet Search Tools ......................................................... 21
- Summary ......................................................................... 27
## 03 DEVELOPING SEARCH STRATEGIES

- Objectives ........................................................................................................................... 28
- Competency Standards ....................................................................................................... 28
- Introduction .................................................................................................................... 29
- Lycos .................................................................................................................................... 30
- WebCrawler ..................................................................................................................... 31
- EINet Galaxy .................................................................................................................... 33
- InfoSeek ............................................................................................................................. 34
- Summary ............................................................................................................................ 35
- References ....................................................................................................................... 36
- Appendix 1 ....................................................................................................................... 38
- Appendix 2 ....................................................................................................................... 39
- Appendix 3 ....................................................................................................................... 45
INTRODUCTION

This Unit is designed to provide information professionals with some basic search skills in searching the indexes of the World Wide Web (WWW). The WWW can accessed through the public Internet. WWW’s resources are organised to provide easy access to users by means of formatted texts, sound, images and other links. This unit will concentrate on some of the better developed search engines of the WWW. It will enable the participants to design Boolean logic-based search strategies to retrieve required information.

The World Wide Web, was put together through a series of protocols developed by Tim Berners-Lee at the European Organisation for Nuclear Research (CERN) near Geneva, Switzerland, during the late 1980s. The WWW’s browsers are constantly changing and evolving. New on-line servers are continually being added to the WWW and its many search mechanisms are always being refined. This unit will be limited to selected search mechanisms of the WWW, due to the vastness of its resources.

This Unit consists of three Modules. Students seeking accreditation are required to complete the assignment (Appendix 3) and submit it for assessment.

Unit Outcomes

By the end of this Unit you should:

- be able to apply the concepts of Boolean logical operators in constructing a search strategy in the World Wide Web environment;
- be able to determine which elements in the search strategy are essential key concepts;
- be familiar with the various approaches used in developing a search strategy using WWW’s search engines;
- understand the concepts of precision and recall;
- appreciate the legal and ethical requirement of searching the vast resources of the World Wide Web; and
- apply evaluation techniques to assessing the outcome of a search.
Library Competency Standards

This course addresses mainly the competencies explained in Unit 49 (Level 6) *Provide clients with specialist and complex access to information* and Unit 17 (Level 4) *Contribute to client access to information* of the Library Competency Standards developed by Arts Training Australia. However, competencies developed by Unit Three: World Wide Web search strategy development skills and knowledge are far reaching. Aspects of the following competency units are affected to some extent:

- Unit 16 (Level 4) Obtain material from remote sources for clients
- Unit 27 (Level 4) Provide research assistance
- Unit 38 (Level 5) Improve accessibility of information
- Unit 45 (Level 5) Collect, analyse and evaluate information for research
- Unit 57 (Level 6) Provide specialist technological support for information access; and
- Unit 65 (Level 7) Manage information access.

Legal and Ethical Dimensions of Search Strategy Development Over the World Wide Web

It almost goes without saying that search strategy development in the electronic information retrieval environment of the Internet has to take place within the parameters of copyright and intellectual property legislation. Although these two very important dimensions fall outside of the immediate instructional focus of this Unit, it is useful to point out that all search strategy development activities for this course should be conducted within the legal and ethical boundaries of accessing the Internet’s information resources.
Unit 3
Upgrading Search Strategy Skills

01 WHAT IS A SEARCH STRATEGY?

Do not complete this Module if you have successfully completed Unit One in this course. Please begin this Unit with Module Two.

Objectives

By the end of this module you should be able to:

- define a search strategy;
- identify the essential decision points in formulating a search strategy;
- understand the basic components of a search strategy;
- identify key concepts in an information request; and
- apply Boolean logical operators to combine the key concepts of a search statement.

Competency Standards

This part of Unit Three provides background information for some of the issues raised by units 49, 21 and 17 of the Library Competency Standards developed by Arts Training Australia.

Introduction

Information retrieval systems are complex systems requiring specialised training. The process of translating an information request into the language of a computer in an on-line, CD-ROM or network environment is called developing a search strategy. Information retrieval systems operate on the principle of a set of matching functions taking place at high speed. The skill and knowledge of the searcher are the main determinants in obtaining the best results from an information retrieval system. A good searcher should have a variety of competencies. These fall into three main categories:
1. General principles of searching skills such as: a general knowledge of information retrieval; the nature of the WWW's resources; and, problem solving, search strategy formulation skills and evaluation of search results. However, human communication skills relating to reference interviews are not very essential for this Unit as resources of the WWW are usually searched by end-users, not by intermediaries. (Unit One: Developing Search Strategy Skills for On-line Searching contains a section on strategies for on-line search intermediaries).

2. System related skills - such as awareness of search software protocols, correct entry forms of search terms and establishing logical combinations on the system.

3. Technical skills in the use of computers.

The Search Process

Regardless of the information retrieval environment - on-line, CD-ROM or World Wide Web - conducting a search contains several stages such as clarifying the search request, deciding on information sources, translating the information request into the language of database systems by identifying its key concepts, conducting the search and evaluating the results. In this Module the emphasis will be on two aspects of information retrieval: (1) the translation of an information request into the language of on-line database systems and (2) the Boolean logic.

Translating the Information Request - Identification of Key Concepts

The search strategy performs two important tasks: it explicitly identifies the concepts in an information request and it also explicitly states the relationships between those concepts. The latter task is made possible with Boolean logical operators.

The first step is to identify correctly the concepts in an information request. For example the statement of “I would like to find some information on the New Zealand Cricket Academy” is a simple search request containing only one concept: the New Zealand Cricket Academy. No other concepts are implied. Whereas, a request of “I am looking for information on breeding alligators” contains two concepts. Breeding and alligators need to appear in the same
article. This is an AND relationship. An information request for "Articles on company intelligence or corporate intelligence" presents an OR relationship between concepts. Articles either on company intelligence or corporate intelligence will satisfy this search request.

Traditionally a standard search strategy form is used by searchers in developing a search strategy. All of the essential components of such a form are provided in Appendix 1. (This form can be copied to make additional forms, as a new form is necessary for each search request). Use this form to develop your search strategies.

ACTIVITY 1

Indicate the number of concepts in the following information requests:

1. Find some articles on global warming caused by methane and chlorofluorocarbons released into the atmosphere.

2. Articles on freedom of speech and religious toleration.

3. Does the use of fossil fuels and clearing of tropical forests contribute to the greenhouse effect?

4. Can you find any articles on censorship and the Internet?

5. Find some recent information on virtual reality.
Boolean Logical Connectors

Search strategy construction in the on-line, CD-ROM or the World Wide Web environment depends on Boolean logical operators. Boolean logic using AND, OR and NOT operators were first introduced by George Boole, the mathematician in 1847. Ironically, Boolean logic remains the main information retrieval tool since its introduction nearly one hundred and fifty years ago, in spite of advancements in information retrieval science. It represents symbolic relationships between concepts or sets. Relationships between concepts can also be viewed in pictorial form using Venn diagrams (See Figures 1 and 2).

Various WWW indices and search engines are being developed to improve its management. Many of the WWW’s search engines are aspiring to offer Boolean searching such as Yahoo, Lycos, WebCrawler, InfoSeek, EINet Galaxy and Veronica/Jughead. Ironically, a better search mechanism than Boolean logic has not been offered yet in this sophisticated hypertext/hypermedia environment.

**AND** logical operator, or connector, requires that the identified search terms that it connects must be present in all documents. In other words, every keyword connected by an AND logical operator must be present for a record to qualify.

**OR** logical operator, or connector, will retrieve documents containing either one or both of the search terms it connects. In other words, the occurrence of any one of the keywords connected by an OR logical operator will qualify the record to be retrieved.

**NOT** logical operator, or connector, will eliminate any documents containing keywords connected by the operator.
Examples of Boolean logical connectors:

"I am looking for information on the World Wide Web and the Hypertext Markup Language" search statement has two concepts: World Wide Web and the Hypertext Markup Language as indicated in the following Venn diagram:

world wide web AND hypertext markup language

![Venn Diagram](image)

Figure 1

The shaded area of the Venn diagram represents items retrieved using the logical connector AND.

"Articles on Internet or World Wide Web" search statement is represented in Figure 2 below:

Internet OR World Wide Web

The shaded area indicates the articles retrieved by this search statement.

![Venn Diagram](image)

Figure 2

**Nesting of Boolean Logical Operators**

Nesting is the term used for the facility that allows a searcher to use several Boolean operators in the same sentence without creating any confusion. Nesting is possible by means of parentheses. Parentheses help to indicate the
Developing search strategies for the World wide Web

order of execution of logical connectors by the system. It is then possible to separate ORs from ANDs:

\[(a \ OR \ b) \ AND \ c\]

In this expression either terms a or b or both will be ANDed with the term c. Whereas this expression will be interpreted differently if it is expressed without the parentheses, thus giving a different result:

\[a \ OR \ b \ AND \ c\]

First, terms b and c will be ANDed as ANDs are normally executed before ORs. The database will then retrieve records containing either a or the intersection of b and c; or both a and the intersection of b and c depending on the use of parentheses.

Consider these two expressions:

\[(\text{sun} \ OR \ \text{solar}) \ AND \ \text{energy}\]

and

\[\text{sun} \ OR \ \text{solar} \ AND \ \text{energy}\]

In the first expression, records containing the concepts of sun or solar energy will be retrieved. The second expression is not correct as the computer will combine solar and energy, then it will retrieve records with only the sun concept.

Stopwords

In a retrieval system, although most of the words occurring in records are indexed, some very commonly occurring words are not. These are known as stopwords. Normally in WWW indexes stopwords are not searched.
Precision and Recall

The objectives of each search are different. Some search objectives call for a narrow search with very few but highly relevant records. Others will require high number of records. In search strategy terminology this dichotomy is expressed as high precision versus high recall. High precision or high recall can be achieved by narrowing or broadening the search parameters. Maximising the retrieval of relevant records while minimising retrieval of irrelevant records is described as maximising both recall and precision. However, there is often a negative correlation between achieving these objectives. In other words, if the precision is very high the recall will be lower; alternatively, if the recall is very high the precision will tend to be lower. The formulae below illustrate the ratios between recall and precision:

\[
\text{Recall ratio} = \frac{\text{Relevant records retrieved}}{\text{Total relevant records in database}}
\]

\[
\text{Precision ratio} = \frac{\text{Relevant records retrieved}}{\text{Total records retrieved}}
\]

Strategies for Narrowing and Broadening a Search

Strategies for narrowing a search include the use of:

- More AND logical operators
- More proximity operators such as \textit{near} between search concepts
- Specific terminology

Strategies for broadening a search include the use of:

- Truncation
- More OR logical operators and fewer AND operators
- Broader terms
- Only the most essential concepts, i.e., fewer concepts
Developing search strategies for the World wide Web

Unit 3

For example:

If too many articles are retrieved by the statement below:

> search and strategy

try narrowing your search by placing a proximity operator such as adj (meaning adjacent) between these two terms. The concept of adjacency will be more fully developed below.

> search adj strategy

On the other hand, if too few articles are retrieved by the search statement:

> information adj literacy

try broadening your search by adding similar concepts:

> (information OR computer) AND literacy

**Truncation**

Computers have the capacity to mask the stem or a portion of a word, to retrieve any character(s) in that masked space. This is called truncation. Most of the WWW's retrieval systems provide a truncation feature; unfortunately truncation symbols vary in different WWW indexing systems. Truncation is a short cut expression of OR logic as it combines a variety of terms, any of which is acceptable, without spelling all the variations of it. The most common type of truncation is applied to the stem of a word.

For example genetic* as a truncated expression will retrieve the following:

> genetic
> genetical
> genetically
> geneticist
> geneticists
> genetics
Proximity or Contextual Operators

The facility to nominate positioning of the words and phrases in records is an added advantage in developing a search strategy. Contextual or proximity operators position terms or concepts in records. If you are trying to locate articles on information flow, the AND logical will not retrieve the most relevant articles. What is needed is a proximity or a contextual operator to indicate that these two terms should be located next to each other in a prescribed order. Tools facilitating this operation are called “proximity operators”. The exact terminology used depends on the WWW indexing system used. For example, Veronica uses adj (adjacent) proximity operator.

\[
\text{information adj flow}
\]

will retrieve these words when they occur adjacent to each other, in this specified order:

\[
\text{information flow}
\]

On the other hand, the near proximity operator will retrieve

\[
\text{information near flow}
\]

both:

\[
\text{flow of information}
\]

and:

\[
\text{information flow}
\]

Here, the proximity operator near searches for the concepts information and flow occurring near to each other in any order. Some indexing systems may not state how many intervening words may come between these two terms. An intervening words facility is especially useful in dealing with stopwords such as in: management by objectives and wash and wear. In these examples, by (a stopword) and and (a Boolean logical connector) are masked.
ACTIVITY 2

Express the following statements by shading appropriate sections of Venn diagrams:

1. \( a \text{ OR } b \)

2. \((a \text{ AND } b) \text{ NOT } c\)

3. \((a \text{ OR } b) \text{ AND } c\)
**ACTIVITY 3**

Express the following Venn diagrams as search statements using $a, b, c$ or $d$.

Please note that the retrieved sections are the shaded portions.

1.

2.

3.
ACTIVITY 4

Express the search statements below by using Venn diagrams.

1. Find an article about the Internet's e-mail services.

2. Locate articles on plant genetics but exclude any references to wheat genetics.

3. Are there any articles written about the greenhouse effect in Australia?

4. Find articles on tropical deforestation or rain forest deforestation in Australia.

5. Find some recent information on video games.

Summary

This Unit establishes the basics of Boolean searching. It grounds two essential aspects of electronic database searching: Boolean logic and translating an information request to the language of electronic databases. These two dimensions of searching constitutes the heart of retrieving information from the World Wide Web.
02 WHAT IS A WORLD WIDE WEB SEARCH?

Developing Search Strategies for Veronica & Jughead, Yahoo and the AliWeb

Objectives

By the end of this module you should be able to:

• understand the basic concepts involved in developing search strategies for Veronica, Jughead and two manual indexes of the WWW: Yahoo and the AliWeb;

• identify the key concepts in a search query;

• be familiar with strategies used for narrowing and broadening a search; and

• understand the concepts of precision and recall.

Competency Standards

This course addresses sections of competencies explained in Unit 49 (Level 6) Provide clients with specialist and complex access to information, Unit 28 (Level 5) Provide clients with access to required information, Unit 21 (Level 4) Maintain accessibility of information and Unit 17 (Level 4) Contribute to client access to information of the Library Competency Standards developed by Arts Training Australia.
### Introduction

As discussed in the introduction to this Unit, the Internet and the World Wide Web are constantly evolving and changing. This Module is limited to selected search mechanisms of the World Wide Web, due to the vastness of Internet resources.

Various Web indices and search engines are being developed to improve its management. Many of the WWW search engines are aspiring to offer Boolean searching such as Yahoo, Lycos, WebCrawler, InfoSeek and EINet Galaxy. So far the only challenge to Boolean searching is presented by WAIS (Wide Area Information Servers) with its “Relevance Feedback” search technique.
Telstra's CUSI as an Organising Facility for the Web

As an organising facility it was decided to select search engines from some of the categories offered by Telstra's CUSI interface to man search engines on the Web. (http://www.telstra.com.au/index/cusi/cusi.html)

1. Configurable Unified Search Interface (CUSI) provides seven categories: Manual WWW indices, such as Yahoo, and CUI W3 Catalog;
2. Robot generated WWW indices, such as Lycos, WebCrawler and InfoSeek;
3. Other Internet indices, such as Veronica and the WAIS Directory of servers;
4. Software indices, such as ArchiePlex and NEXORS Mac Software Catalog;
5. People indices, such as the NetFind Gopher and UFN Search;
6. Document indices, such as RFC Index Search and the CIA World Factbook;
7. Dictionaries, such as Roget's Thesaurus, Webster's and Acronym Dictionary.

The categories of Telstra's CUSI categories are rather curious. Its manual indexes category includes indexes that are generated by human indexers. Yet, even a manually maintained indexing system such as Yahoo is starting to rely more and more on search engines. It was reported in December 1995 that "Yahoo! complemented its 10 human indexers with a spider. The spider was a piece of software licensed from Open Text, a Canadian company that originally developed the software to navigate the database of the Oxford English Dictionary (Ross and Hutheesing, 1995: 68).

The robot generated indexes of the WWW depend more heavily on search engines such as WebCrawler and Lycos to navigate larger chunks of the Internet. For example, Lycos with its software programme called "Longlegs" is able to maintain a catalogue listing 5.6 million web pages. The "Other" category contains Veronica and the WAIS Directory of Servers which are not WWW based. CUSI's last four categories Software, People, Documents and Dictionaries are like a subject grouping of the Internet's resources which are searchable via Telstra's CUSI.

The WWW's Metasearch search engine offers, like CUSI and AltaVista an organising device for WWW's various browsers. Search terms and Boolean logical operators can be entered only once in Metasearch's form to be searched simultaneously in the WWW's several search engines - such as Yahoo, InfoSeek, Lycos, OpenText, etc.

Due to the large number of search engines available over the WWW and the Internet only six were chosen from CUSI's first three categories.
In Unit Two of this Module Veronica/Jughead (from Other Internet indices category) as well as Yahoo and AliWeb (from the Manual indices category of CUSI) will be examined. Unit Three will continue with four robot generated WWW indices: Lycos, WebCrawler, EINet Galaxy and InfoSeek.

**Veronica (Very Easy Rodent-Oriented Net-Wide Index to Computerised Archives) and Jughead (Jonzy’s Universal Gopher Hierarchy Excavation and Display) - Two Internet Search Tools**

Veronica was created, at the University of Nevada-Reno, as a search tool for retrieving resources from Gopher servers all over the world. Veronica acts like an index to titles and directory names on the menus for most Gopher sites registered with the central Gopher server at the University of Minnesota. It acts as a search engine for approximately 6000 Gopher servers, containing over ten million items. It is an index as well as a retrieval system. Veronica searches keywords in titles as they appear on the menu of its Home Gopher server; it cannot do a full-text search of the contents of the resources.

You should ...think of Veronica in somewhat the same terms you think of archie. Veronica must collect data from the Gopher servers it consults and then make searches of those databases available to the Gopher clients that access it. .. Veronica is not a standalone program; there are no Veronica client programs in the same sense that Gopher clients have proliferated on the network. Rather, Veronica remains a cohesive link to Gopher, working with the Gopher client of your choice to present the results of its searches and make their connections available to you. (Gilster, 1994: 79)

**Veronica and Jughead Boolean Searching**

Veronica facilitates Boolean searching with all three operators AND, OR and NOT as well as truncation of search terms.

If you use a simple multiple-word query, it is the same as using AND between the words. For instance ‘acid rain’ is the same query as ‘rain and acid’ ...we recommend using AND to create a tightly-focused query. We recommend that the word OR be used very rarely. Usually, OR just produces thousands of hit-or-miss results. OR is best used in conjunction with other operators, as ‘rice and (fried or curr*)'.

Many other veronica-access menus offer a single entry rather than a list of servers. In this case, simple click on the search type desired, and submit your query in the dialog (sic) box. ... Most access menus offer two predefined search types:

> Search GopherSpace by keywords in Titles

This search will find ALL TYPES of resources whose titles contain your specified search words. The resources may be of any Gopher data TYPE; e.g. ascii documents, gopher directories, image files, binary files, etc.

> Search Gopher DIRECTORIES ONLY for keywords in Titles.

This search will find only Gopher directories whose titles contain the specified words. (Foster, 20 November 1994)

This second option is useful for identifying only major Gopher directories for making the search more manageable.

In addition, certain flags help Veronica to locate resources of a specific type. The following is a summary of the options. (It should be pointed out that not all Veronica servers support the (-l) linking option):

- t limits the search to items of specified data type(s)
- m specifies maximum number of hits
- l creates a file of links for the retrieved resources

**Problems with Veronica**

The main problem with Veronica is its lack of uniformity. Gopher menus do not adhere to any universal standards. Duplication occurs, as some files exist at different Gopher sites sometimes under different names. More importantly, Veronica is extremely difficult to access due to high user traffic. Veronica searches through previously retrieved indexes, therefore it does not conduct a real-time search. In spite of these problems Veronica is suggested as a searcher's first stop (Gilster, 1995).
Jughead

Jughead is a subset of Veronica; it performs Boolean and keyword searches of one particular Gopher server. It was created at the University Utah by Rhett Jones. Jughead was designed to restrict searches of “gopherspace” as much as possible by limiting the search usually to one Gopher server. “The system administrator at a university can set up a Jughead server to allow users to search only the Gopher at that university” (Gilster, 1995: 427). Examine Edith Cowan University’s Gopher which is located at: Distributed Document Service <GOPHER_ECU> (Root gopher server: scorpion.cowan.edu.au).

For example the University of Texas at Austin sets its Jughead on its Gopher as follows:

Jughead: Search gopher menus at UT Austin

--> 1. About. jughead.
   2. Jughead: Search menus in UT Austin gopherspace <?>

ACTIVITY 5

Connect to:

  gopher://veronica.scs.unr.edu/1/veronica
  or
  gopher://gopher.scs.unr.edu:70/11/veronica

Via Telnet or the World Wide Web. (Access via WWW provides a more user friendly approach).

- Select from the main menu one of the Veronica server sites.
- Select a menu item that begins with the words: {Search GopherSpace by Title word(s)........}
- Type a search statement such as:

  lawyer* and marketing not advertise

  (N.B.: Boolean logical connectors are highlighted here for the sake of clarity. Do not highlight them in your searches.)
ACTIVITY 6

Conduct several searches using Boolean connectors (and, or, not) and truncation.

Be very patient and persistent as Veronica's servers are usually very busy!

Comment on your search experience with Veronica (or Jughead).

Yahoo
(http://www.yahoo.com/)
&
(http://www.yahoo.com/search.html)

Yahoo as a topic oriented catalogue of the Internet sites offers Boolean searching to find matches in title, URL sites and Comments fields. Logical combinations AND and OR are supported.

Yahoo’s Search Limiting Facilities

Yahoo offers a topic tree containing the following categories:

- Arts
- Business and economy
- Computers and Internet
- Education
- Entertainment
- News
- Recreation and sports
- Reference
- Regional
- Science
Developing search strategies for the World wide Web

Unit 3

Limiting by the date of new listings option limits a search to listings added on the current day; during the week, during the last month and during the last three years. However, the year default limit can be altered by the user.

The Truncation issue is addressed by the choices of "substrings" and "whole words". For example, a substring search for engine in:

```
search and engine substring ✓
```

statement retrieved records on search engines as well. Thus the substring facility acts as a truncation feature.

ACTIVITY 7

Find the on-line services provided by NASA using Yahoo's Boolean logical operator AND.

AliWeb

http://web.nexor.co.uk/public/aliweb/doc/search.html

AliWeb, a public service provided by NEXOR, Ltd., in Nottingham, U.K., offers Boolean searching with AND and OR logical connectors. In addition, AliWeb provides truncation under a "substring" option with three choices: substring; whole word or regular expression.

AliWeb comes closest to a bibliographic database searching by its field searching limits by:

- title
- descriptor
- keywords
- URLs
- other fields.

In addition it has a limiting function by country under "You can restrict the results to a domain (e.g., "U.K.") option."
AliWeb's other value-added feature is its output choices which users can select in addition to titles (the default choice) these are:

- Descriptors
- Keywords
- URLs
- Other fields.

**Case Sensitive Searching**

AliWeb's case sensitive searching facility is very useful in achieving precision. For example, it can discriminate AIDS from aids, thus eliminating unwanted noise in a search.

One main disadvantage of AliWeb was its availability only from a single site. However, this problem has been solved now. There are various mirrors (sites) of the AliWeb database, including one in Australia, located at: http://www.webnet.com.au/aliweb/

**ACTIVITY 8**

Several highly relevant references are required on AIDS research done in the United Kingdom.

Choose AliWeb's Case Sensitive option; truncate appropriate terms and browse retrieved items in full (record) format.

**ACTIVITY 9**

Design a search strategy that will require the use of the logical connector OR.
Summary

Initially two of the Internet's search tools, Veronica and Jughead were examined to set the scene for World Wide World's hypertext/hypermedia search browsers. This newer medium with links across the global Internet provides on-line searchers with easier access to large scale distributed multimedia resources that can be searched with fairly sophisticated search tools - including Boolean logical connectors, proximity operators, truncation and controlled output options. In this Module, two such browsers were studied: Yahoo and AliWeb.
Developing Search Strategies:
Lycos, WebCrawler, EINet Galaxy & InfoSeek

Objectives

By the end of this module you should be able to:

• appreciate information retrieval facilities of four of the search engines of the World Wide Web:
  - Lycos
  - WebCrawler
  - EINet Galaxy
  - InfoSeek;

• perform basic search activities taking advantage of the various information retrieval tools of these indices of the WWW; and

• appreciate the strengths and shortcomings of the information retrieval capabilities of these four robot generated WWW indices.

Competency Standards

This part of Module One provides background information for issues raised by units 49, 21 and 17 of the Library Competency Standards developed by Arts Training Australia.
Developing search strategies for the World wide Web

Unit 17
Element Performance Criteria
4 1,2,3

Unit 21
Element Performance Criteria
1 1
2 4,5

Unit 49
Element Performance Criteria
2 4
3 6

Introduction

In this Unit the examination of World Wide Web browsers are continued with: Lycos, one of the largest robot maintained browsers of the WWW; WebCrawler, a smaller but perhaps more up-to-date index to WWW's resources and two others, EiNet Galaxy and InfoSeek.
Lycos


Lycos, one of the largest Robot indices of the Web, supports Boolean searching with the aid of pull-down menus. Lycos means spider in Latin. It provides a subject tree with the following subject headings:

- Arts/Humanities
- Bus./Finance
- Computers
- Education
- Government
- Health/Medicine
- Internet
- Entertainment
- Just for Kids
- News/Info
- Science/Tech.
- Shopping
- Social Issues
- Sports
- Road Less travelled
- The World

Lycos supports Boolean searching with the aid of pull-down menus. For example, "If you type 'jeep cherokee' as your query, Lycos will find all documents containing either 'jeep' OR 'cherokee'" "This is the 'match any term (OR)' search option... [AND search option is provided by the 'match all terms (AND)' pull-down menu choice.""] (Lycos, 20 November 1995). “matching up-to seven terms” ‘Option’ is useful in matching terms that may have different spellings, such as labor or labour. The “Loose match” search options box serves as a type of precision tool.

When set to ‘loose match’ you will get more documents, but they will tend to be less relevant to the query you made ... If you want the Lycos search engine to be more selective, change the search Option ... to ‘strict match’. Lycos will return only documents which have a very high relevance to your query. (Lycos, 20 November 1995)
Lastly, Lycos offers standard (the default), detailed (the fullest record) or minimum (summary) display options for displaying the results which are shown 10-40 per page.

**ACTIVITY 10**

Locate some information on computer games for children.

**WebCrawler**

http://www.webcrawler.com/

WebCrawler, is a smaller but a more up-to-date robot generated index, according to Telstra's CUSI. It offers AND and OR Boolean logical operators with its pull down menu options of: “Find pages with all or any of these words.”

WebCrawler claims that the system “understands Plain English and is programmed with novice users in mind” (WebCrawler, 20 June 1996). The Novice search form contains a dialog box and an output option for titles or summaries. Results can be viewed in short (just a list of titles) or detailed format (a list of titles plus summaries).

Ranking of Results:

WebCrawler provides a “ranking of results” feature as explained below: Ranking of Results - The results are provided in order of relevance with the most relevant resources first. The icons to the left of the results provide a graphical indication of how relevant each link is to your search, being the most relevant, and the least. A detailed explanation of how these numbers are calculated is provided on the page of Frequently Asked Questions (FAQ).

(http://webcrawler.com/WebCrawler/Help/Results.html, 8 July 1996)

WebCrawler gives the following search hint for broadening a search: “If your search on bed and breakfast in Northern California produced too few relevant resources, try bed and breakfasts inns “small hotels” in Northern California (http://webcrawler.com/WebCrawler/Help/ImproveResults.html, 8 July 1996).

Users experienced with Boolean searching are advised to skip to the section on Advanced searching.
Help in the “Advanced Searching” section provides a useful summary to the information retrieval facilities of WebCrawler as shown in the following (abbreviated) segment.

<table>
<thead>
<tr>
<th>Operator</th>
<th>Example</th>
<th>Finds</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>gardening AND vegetables</td>
<td>pages containing both gardening and vegetables</td>
</tr>
<tr>
<td>OR</td>
<td>whales OR cetaceans</td>
<td>pages containing either / or / both of whales or cetaceans</td>
</tr>
<tr>
<td>NOT</td>
<td>science NOT fiction</td>
<td>pages containing science but not fiction retrieved</td>
</tr>
<tr>
<td>NEAR</td>
<td>arthritis NEAR /25</td>
<td>pages both terms within 25 words of each - in any order will be retrieved</td>
</tr>
<tr>
<td>ADJ</td>
<td>global ADJ warming</td>
<td>pages with these terms will be retrieved</td>
</tr>
<tr>
<td>&quot;.........&quot;</td>
<td>&quot;all you can eat&quot;</td>
<td>pages with this phrase will be retrieved - regardless of a stop work such as all</td>
</tr>
<tr>
<td>(.........)</td>
<td>Homer NOT (Simpson OR Alaska)</td>
<td>pages with Homer - but not if appears with Simpson OR Alaska</td>
</tr>
</tbody>
</table>

(WebCrawler, 20 June 1996)
EINet Galaxy

http://galaxy.einet.net/

EINet Galaxy is provided as a public service by TradeWave Corporation. EINet Galaxy, like Yahoo gives a subject tree with following headings:

- Business and Commerce
- Community
- Engineering and Technology
- Government
- Humanities
- Law
- Leisure and Recreation
- Medicine
- Reference
- Science
- Social Sciences

Its Boolean search engine, like many others, makes available “Match any search term, or all search terms” options. Its other facilities include: “all text” or “title text” searching. All text search choice facilitates information retrieval beyond titles. On the search form, links are provided to Gopher titles or Telnet resources. Its one interesting feature is the link text function. The output of a search can be controlled by the user - such as the long, medium and short outputs. It also allows the user several choices such as opening a retrieved site or document, adding a bookmark or opening a window to the retrieved source.

ACTIVITY II

Locate some recent information about the next Olympics. Conduct this search first in the “Sports” then in the Leisure & Recreation subjects.
InfoSeek

http://www.infoseek.com/

This search engine offers some interesting features such as capitalised proper names and word adjacency. It contains a subject tree as shown below:

- Arts & Entertainment
- Business & Finance
- Computers & Internet
- Education
- Government & Politics
- Health & Medicine
- Living
- News
- Reference
- Science & Technology
- Sports
- Travel

It provides useful help information on searching the Web:

Capitalize names of people and places, and use a comma to separate lists of names.

To identify a group of words that must appear together in order, use double quotation marks around them, or hyphenate them.
(InfoSeek, 20 June, 1996)

Don’t attempt to use words like and or or as logical operators...[as] The software handles all letters as search terms. When you type common words like and or or lowercase, the software doesn’t search for them because they are very common words. When you type such words uppercase, the software attempts to find uppercase instances of the words.
(InfoSeek, 20 November, 1995)

InfoSeek’s Syntax section gives very detailed information about its search engine with rules and examples:
### Example

| Clinton, Gore | Use a comma between unrelated capitalized proper names. |
| "user interface" | Use double quotation marks around words that should appear next to each other. |
| user-interface | Use a hyphen between words that should appear very close to each other. |
| [WWW search] | Use brackets around words that should appear near each other in any order. |
| chip +Pentium | Use a plus sign in front of a required word or phrase. Use a minus sign in front of word or phrase you don’t want to appear anywhere in the resulting documents. |

(InfoSeek, 20 November, 1995)

Thus, it can be said that the minus (-) sign substitutes for the NOT Boolean operation. Word proximity is provided by means of brackets and hyphens; double quotation marks facilitate string searching.

### ACTIVITY 12

Conduct several searches following the instructions given in the Rules section above.

### Summary

Unit Three continues with the examination of the World Wide Web’s better known browsers by investigating four more search engines: Lycos, WebCrawler, EINet Galaxy and InfoSeek.
REFERENCES

EINet (20 November 1995). http://galaxy.einet.net/cgi-bin/wais-text-multi


Lycos (20 November, 1995). http://lycos.cs.cmu.edu/


http://www.wais.com/newhomepages/surf.html

This Unit does not contain a bibliography as any printed documents would be out of date as soon as they are published. One of the best ways to keep up with the WWW is simply to use it regularly.
## Appendix 1

### World Wide World Search Strategy Form

Search request (expressed in a sentence):

<table>
<thead>
<tr>
<th>Concept One</th>
<th>Concept Two</th>
<th>Concept Three</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search strategy
Appendix 2

Answers

01 WHAT IS A SEARCH STRATEGY?

ACTIVITY 1

Indicate the number of concepts in the following information requests:

1. Find some articles on global warming caused by methane and chlorofluorocarbons released into the atmosphere.

Concept 1: global warming, greenhouse effect, ozone depletion, etc.
Concept 2: chlorofluorocarbons etc.

2. Articles on freedom of speech and religious toleration.

Concept 1: freedom of speech etc.
Concept 2: religious toleration etc.

3. Does the use of fossil fuels and clearing of tropical forests contribute to the greenhouse effect?

Concept 1: fossil fuels, etc.
Concept 2: tropical forests, rain forests, etc.
Concept 3: greenhouse effect, ozone depletion, etc.
4. Can you find any articles on censorship and the Internet?

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>censorship etc.</td>
<td>internet</td>
<td>world wide web</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>www</td>
</tr>
<tr>
<td></td>
<td></td>
<td>etc.</td>
</tr>
</tbody>
</table>

5. Find some recent information on virtual reality.

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>virtual</td>
<td>reality</td>
<td></td>
</tr>
</tbody>
</table>

**ACTIVITY 2**

Express the following statements by shading appropriate sections of Venn diagrams:

1. \( a \ OR \ b \)
2. (a AND b) NOT c

3. (a OR b) AND c
ACTIVITY 3

Express the following Venn diagrams as search statements using a, b, c or d.

Please note that retrieved sections are the shaded portions.

1. \((b \text{ OR } c) \text{ NOT } (a \text{ OR } d)\)

2. \(a \text{ AND } b \text{ AND } c\)

3. \(a \text{ NOT } (b \text{ OR } c)\)
ACTIVITY 4

Express search statements below by using Venn diagrams.

1. Find an article about Internet’s e-mail services.

2. Locate articles on plant genetics but exclude any references to the wheat genetics.

3. Are there any articles written about the greenhouse effect in Australia?
4. Find articles on tropical deforestation or rain forest deforestation in Australia.

5. Find some recent information on video games.

ACTIVITIES 5 - 12

Answers to these activities will depend upon the search which is undertaken. Activities in Modules 2 and 3 can contribute to the assessable assignment described in Appendix 3.
Appendix 3

Assessable Assignment

The assessable assignment for Unit Three: Developing Search Strategy Skills for Searching the World Wide Web has two sections: hands-on tutorial exercises and the evaluation of these exercises.

1. Submit printed results of any three of the hands-on tutorial exercises provided in Module 2 or 3 of this Unit.

2. Accompany the search results with an evaluation of these searches in the context of information retrieval facilities of the World Wide Web.

The length of the evaluation section should be approximately 500 words.