Chapter 5 Access of Information

In this chapter, you will learn:

1. The concepts and usage of some communication and information access tools on the Internet.
2. To use search engines and subject directories on the Internet and to perform effective information searching and discovery using these tools.
1. Information Search Strategies

(1) Pre-search Analysis:
- Identify any societies, organizations that have the information you sought at their websites. Go to the websites to look for any pointers or information.
- Identify any distinctive words, phrases, acronyms associated with the topic.
- Identify other words that are likely to appear in any webpages over the topic.
- Identify any synonyms, variations in spelling for the previously identified words or phrases: the OR search is useful here.
- Identify any irrelevant documents these search words/phrases may pick up. What are the words that distinguish these irrelevant documents? We can see that the NOT search is useful here.
- Identify other words/phrases to describe the broader subject area that may be useful when searching a subject directory e.g. communicable disease, infectious disease as opposed to SARS.
- Prepare the search terms according to the techniques discussed in Boolean Search Term Techniques. That is, use long phrases, use OR for synonyms or spelling variations, and use NOT for eliminating possible irrelevant pages.

(2) Executing the Search: Use your prepared search terms in a search engine.
If your search returns too many matches:
- Add more AND terms to pinpoint your area of interest.
- Add more NOT terms to eliminate irrelevant matches.
- Some search engines allow limiting matches by specifying date of publication, languages, webpage URL, etc.

If your search returns too little matches:
- Try reducing the number of AND terms to broaden your area of focus
- Try adding variants of your search terms with OR
- Use a broader term describing the subject matter in the search

(3) To Look for an Overview

Search in a subject directory using the broader subject term and follow the links. If there are too many links, search in the subject directory using your narrower keywords. Links from the subject category should point you to the main sites about the subject matter. Sometimes looking up the subject matter in an encyclopedia may help.
(4) Seek Expert Advices

Seek advice from relevant mailing lists, newsgroups, or other discussion groups. Sometimes you come across experts in those forums who can point you to difficult to find articles or resources. Search DejaNews if you think the topic may have been previously discussed on Usenet.

2. World Wide Web (WWW)

Historically we have computerized hypertext systems – electronic documents which contain links to other electronic documents. Hypertext systems may appear in, for example, a CDROM containing an encyclopedia, with one topic linking to other topics. Users can easily move from one topic to related topics through the links. The links themselves are called hyperlinks.

When we applied the concept of a hypertext system to the Internet, we could then have a vast collection of hypertext documents, with hyperlinks between them, using the Internet as the media. The result is the World Wide Web (WWW). In the WWW, multimedia contents can coexist with text. We can say that the WWW is a multimedia hypertext system.

When we access a webpage on the WWW, the webpage can be a fixed document; or it could be a webpage generated instantly and dynamically by computer programs extracting information from a specialized database. For example, the database can be a library catalogue, or a bookstore catalogue, or a news database. Or in the case of a web-based discussion forum, the webpage can be generated from the database containing the individual messages of the discussion forum.

Search engines are used for finding webpages of interest among the vast sea of pages in the WWW. A search engine is a special website that allows users to input keywords. The search engine will do an Internet-wide searches using the keywords to present a list of links to the user, pointing the users to webpages that contains the keywords he inputted.

Search engines do not really "search" the Internet out there when the users entered the keywords. Instead a search engine maintains a large index for a huge numbers of Internet sites by retrieving each individual webpages. Some claim to have indexed over 3 billion webpages. In essence they maintain (sort of) a copy of a portion of the WWW. Contrast this with specialized search sites such as library catalogs, DejaNews, which only have very limited domain of information. Popular Search Engines on the WWW include:
Yahoo have a list of Internet search engines at:
http://dir.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Search_Engines_and_Directories/

We remembered that a webpage can be fixed or can be generated by computer programs. Pages that are dynamically generated from a database may or may not be searchable in a search engine. Most likely search engines would not have visited the page and have an index on it. In this sense, many store catalogue, online dictionaries, library catalogue, etc cannot be searchable on search engines. Pages that required a password or payment or pages that are not linked to by anyone may also not be searchable. These portions of the WWW are collectively called the "invisible web".

When we use a search engine, the resulting links are ranked by relevance of the page to the search keywords. A link in the top should indicate that the link is more relevant to the search. Importance of the pages also plays a part. More links to the page from important sites indicates more importance of the page. Important pages also should appear at the top of the search results. Different search engines will use different method to calculate relevance and importance, and this is called the engines’ “page ranking policies”. Page ranking policies affects the usefulness of the search engines to the users.

We have to beware some search engines put advertisers' pages at the top regardless of the pages’ relevance or importance. This practice is called “paid placement”. Nowadays search engines usually labeled (in small prints) these advertisers’ pages as "sponsored links" or "sponsored listing". You should read the headings carefully when evaluating the results from a search engine.

For all of their popularity and importance, search services like Google have a significant limitation: They don’t answer questions or provide information directly. Instead, they will generate a list of Web sites where the answers might – or might not –be found. But now there is an entire search service devoted to providing direct answers to search queries. It is called Answers.com, and it is available at http://www.answers.com. Using a variety of reference sources, such as dictionaries and encyclopedias, it generates a thoughtfully
organized page of relevant information about your search query without requiring you to click any further Web links.

3. **Meta-search Engines**

There are some search sites that do not maintain indexes of webpages. Instead they search several other search engines with the user inputted keywords, then combine and organize the results from those searches to form a result page for the user. These sites are called meta-search engines.

The reasoning is that searching several search engines will increase the information space for the search, and that resources missed by one search engine can be covered by the others. But this idea of meta-searching is often better than the reality it presented, as the main problem for searching is usually too much information instead of not enough. The usefulness of a meta-search engine then in practice depends on how good the engine combines and organizes the large number of links in the results. Meta-search Engines on the Net include:

- Search.com [http://www.search.com](http://www.search.com)

4. **Boolean or Advanced Search**

Most search engines support searches with phrases in addition to searches with keywords, e.g. "world health organization" as a single phrase as opposed to "world" "health" "organization"

Most search engines also allow searches with a combination of keywords connected by AND, OR, NOT. In practice, AND is usually implicit by default by most search engines. That is, “world” “health” “organization” actually means “world” AND “health” AND “organization”. We called the searching with AND, OR, NOT connectives “boolean searches”.

Since each search engines use different syntax and rules for the AND, OR, NOT connectives, you should check up the help pages for your favorite search engines.

OR search example: suppose we want to find pages about museums in London or in Paris:

- museum (london OR paris)
NOT search example: suppose we want to find pages about comparative literature but want to exclude references to university departments:

- "comparative literature" NOT "department of"
  
  Google uses '-' to indicate NOT
- on Google: "comparative literature" -"department of"

Each search engine also has a list of “stop words”. They are common words such as ‘the’ ‘on’ 'and’ ‘’, etc that are ignored by the search engine as search keywords. For example, most search engines will regard a search for:

- catcher in the rye

as a search for

- catcher AND rye

In this particular case of course the phrase search:

- “catcher in the rye”

may best serve the user’s purpose.

A few search engines support "wildcards": e.g. book* will match books, bookstore, bookworms, etc. This and other convenient functions are disappointedly only supported on a few search engines each with their own different rules.

5. Search Term Techniques

A number of techniques is often use to guide us to better utilize the boolean connectives AND, OR, NOT and the phrase searching functions.

- Does your search identify a distinct proper name or phrase?
  
e.g. "lingnan university"  "world health organization"
  "george w. bush" "country road take me home"

Proper names or a long phrase help narrow in the results, reducing the number of matches and making the results more relevant.

- Use more AND terms to narrow the search.
  "president bush" AND "iraq war" AND WMD AND claims

More AND terms, as in a long phrase, narrow in the results and reduce the number of matches.
e.g. lingnan university have 12,200 matches in Google, while lingnan university gymnasium have only 67 matches

• Use OR to specify a set of synonyms or spelling variations. "president bush" OR "george bush" OR "george w. bush" (airline OR airlines) AND profitability olympics AND sarajevo OR sarayevo

Using more OR terms will increase the number of matches. e.g. profitability airline have 66,700 matches in Google, while profitability airline OR airlines have 121,000 matches.

• Use NOT to eliminate groups of matches. "comparative literature" NOT "department of "sars NOT "south african revenue service" NOT "south africa"

Obviously using more NOT terms reduces the number of matches and eliminates irrelevant results.
e.g. sars have 4,450,000 matches in Google, while sars -"south africa" have 2,500,000 matches

6. Subject Directories

Some search sites, e.g. Yahoo http://www.yahoo.com, can be classified as a Subject Directory. Subject Directories provide a searchable hierarchy of subject matters. All linked pages are classified (by human) as belonging to certain subject matters. In this way, users can found webpages related to a subject matter; as opposed to webpages related to a keyword as in a search engine.
Some directories also provide evaluation by human expert to the quality of the linked webpages. Usually the subject directories also provide search engine services.

Popular Subject Directories include:

• Yahoo http://www.yahoo.com/
• Google Directory http://directory.google.com/

Popular Subject Directories that are specialized in academic information include:

• Librarians' Index http://www.lii.org/
• Infomine http://infomine.ucr.edu/
• Academic Info http://www.academicinfo.net/
7. Usenet News

Also known as Network News, Usenet news is not actually news. Instead Usenet is a global system on the Internet consisting of a large number of electronic discussion forums. Each forum is called a “newsgroup”. These newsgroups are not part of the WWW and users should use special software called a newsreader to access them.

Usenet newsgroups (i.e. the forums) are divided into categories. The category of a newsgroup is indicated in the newsgroup’s name. The name is a very concise description of the subject matter the newsgroup is discussing. There are seven main Usenet newsgroups categories. We will show some example newsgroup names within the seven categories below:

- **comp.* category**
  - comp.games.development.design Discuss computer games design for game developers
  - comp.lang.c++ Discuss the programming language C++

- **news.* category**
  - news.admin.net-abuse.misc Discuss abuses on the network systems
  - news.groups.questions Discuss questions about newsgroups formation

- **rec.* category**
  - rec.music.beatles.moderated Discussion the Beatles, with moderator
  - rec.pets.cats Discussion raising cats as pets

- **sci.* category**
  - sci.physics.relativity Discuss the theory of relativity
  - sci.engr.chem. Discuss chemical engineering

- **soc.* category**
  - soc.culture.japan Discuss Japanese society and culture
  - soc.support.depression.family Support group for family with depressed member

- **talk.* category**
  - talk.politics.european-union Discuss the politics of the EU
  - talk.religion.newage Discuss new age religions

- **misc.* category**
  - misc.education.medical Discuss medical education
  - misc.consumers.house Discuss consumer issues regarding home buying
Besides the seven main categories, a large number of other categories exist within Usenet. But they may not have the same global distribution and the large participation that the seven main categories enjoy. They tend to be more specialized in the subject matter and audience. They include:

- **alt.* category**
  - alt.sex Discussions about sex
  - alt.sport.snooker Discuss the game of snooker

alt.* groups tend to have subject matters that are more alternative in taste. But since alt.* groups have less strict creation rules than regular groups, some legitimate topics also appear in alt.* as newsgroup. alt.sport.snooker above is an example.

- **bionet.* category**
  - bionet.toxicology Discuss toxicology. For biologists.
  - bionet.entomology Discuss entomology. For biologists.

- **linux.* category**
  - linux.redhat.misc Discuss the Red Hat Linux operating system
  - linux.samba Discuss the Samba file system

- **netscape.* category**
  - netscape.public.dev.css Discuss the CSS specifications
  - netscape.public.general Discuss general matters about Netscape products

There are also a large number of regional newsgroups. We can expect them to have limited distribution and users restricting to local people only.

- **hk.* category**
  - hk.rec.sport.soccer Discuss soccer.
  - hk.biz.general Discuss businesses.

- **tw.* category**
  - tw.bbs.soc.politics Discuss politics.
  - tw.bbs.literral.poem Discuss poems.

- **lingnan.***
  - lingnan.forum General matters discussions.
  - lingnan.hostelc Discuss matters relating to Hostel C.

Even if your local computer network have installed and participated in the Usenet system, the news administrator of your site could determine what newsgroups its users can see, and how
long the discussion messages are kept. So you may not be able to read all the newsgroups listed above, or may not be able to read messages posted say, a week ago.

Users use a newsreader to read or send (we usually use the word “post”, as in the sense of “posting a message to a notice board”) discussion messages. Popular newsreader includes: Forte Free Agent and Outlook Express.

In newsgroups, a discussion message in a newsgroup is called a post. Posts are grouped into threads. A thread is just the chain of posts that are discussing the same subject.

Since people may have asked the same questions over and over again on a newsgroup, some volunteers may have prepared a list of frequently asked questions and their answers for the newsgroup, to save people’s time asking and answering. These lists are called the FAQs of the newsgroup. An archive collecting all FAQs for all newsgroups exists at http://www.faqs.org/. New users are suggested to read the FAQs of interested newsgroups, to avoid asking the same questions again on the newsgroup.

Many users do not realized that the act of posting on Usenet may leave a permanent record of the user on the Net, since the posts of Usenet are constantly being monitored and archived by some security or commercial organizations. For example, an archive of all previous posts is searchable on http://groups.google.com/, commonly known as the DejaNews archive.

As an information source, Usenet have the same problem as the Internet in general. Some newsgroups are awash with many uninformative or inaccurate posts, posted by anonymous users. But of course the culture of different newsgroups differs, and the usefulness to you should be evaluated in a newsgroup by newsgroup basis.