
UNIT 19 SEARCH STRATEGIES, PROCESSES AND TECHNIQUES

Structure

- 19.0 Objectives
- 19.1 Introduction
- 19.2 Search File – An Essential Component
- 19.3 Search Strategies and Pre-requisites
 - 19.3.1 The Pre-Search Interview
- 19.4 Search Techniques
 - 19.4.1 Keyword and Phrase Search
 - 19.4.2 Keyword and Subject Search
 - 19.4.3 Boolean Search
 - 19.4.4 Truncation Search
 - 19.4.5 Proximity Search
 - 19.4.6 Field-specific Search
 - 19.4.7 Limiting Search
 - 19.4.8 Range Search
- 19.5 The Information Search Process
 - 19.5.1 Basic Features
 - 19.5.2 Search Tactics
- 19.6 Online Searching
 - 19.6.1 Online Search Services
 - 19.6.2 Basic Steps in an Online Search
 - 19.6.3 Features of an Online Search Service: DialogWeb
- 19.7 Access to Information on the Web: The Tools
- 19.8 How the Search Engines Work
 - 19.8.1 The Spider
 - 19.8.2 Search Engine Software
 - 19.8.3 Indexing
- 19.9 Common Search and Retrieval Features of Web Search Engines
 - 19.9.1 Specialised Search Engines
- 19.10 Multiple Database Searching and Common Command Language
- 19.11 Summary
- 19.12 Answers to Self Check Exercises
- 19.13 Keywords
- 19.14 References and Further Reading

19.0 OBJECTIVES

In any ISAR system searching is the most vital component for retrieval of information. It is necessary that effective strategies may be developed and adopted for this purpose. In the searching process there are many techniques available. Adoption of the appropriate technique is dependent upon the users' requirements, and features and facilities available in the system. This Unit will make you conversant with all these aspects.

After reading this Unit, you will be able to:

- 1 learn what is a search strategy;
- 1 understand an information search process;
- 1 learn the techniques used to conduct an online information search;
- 1 get a basic understanding of how the web search engines work;
- 1 get an idea of the information retrieval features of web search engines; and
- 1 successfully conduct information searches using online search services and web search engines.

19.1 INTRODUCTION

You must have understood that the searching subsystem is one of the major subsystems of an information retrieval system. In this subsystem, users' queries are received and interpreted, appropriate search statements are formulated, and the actual search (i.e., matching queries with the surrogates of information resources file) is conducted with a view to retrieving the required information. All these tasks can be performed manually, as is used to be done in the earlier systems, or these can be automated.

The information retrieval may be of different types – reference, document, fact and knowledge. In the reference retrieval, information related to specific questions are retrieved. In document retrieval, retrieval by surrogates of documents such as author, title, subject, etc. are made which may provide the information on a bibliographic citation with or without abstracts. Nowadays complete texts are also retrieved (text retrieval system). In the fact retrieval system, specific data or facts are retrieved (viz., numerical databases). The knowledge retrieval system is a rule based system in which there is a knowledge base with capability for knowledge acquisition and inference engine. It processes facts, information into knowledge, based on rules.

The advancements in information and communication technologies (ICT) brought out the possibility of more dynamic searching via online methods. The concept of online searching has occupied a large and significant area in the study and research of modern information retrieval. Online information retrieval systems appeared over four decades ago, and over the years they have developed significantly in terms of number, features and facilities. However, a user often faces difficulties in approaching an online information retrieval system, especially in formulating an appropriate search statement.

The cost of searching a database, whether in-house or external, can be reduced significantly if an appropriate strategy for searching is followed. The search strategy helps the user select the optimum path for searching a file or a database. This involves a number of measures that are to be taken before and during a search. This Unit covers the basic concepts of the search strategy and describes the actual searching process in the context of online and web information retrieval systems.

19.2 SEARCH FILE – AN ESSENTIAL COMPONENT

A search file is an essential component of a database which describes the document collection of IR system. Generally the search for information is through subjects or keywords. This implies that every aspect or characteristics of documents are to be properly identified and indexed which form the essential component of a search file. The steps in creating a search file are:

- a) deciding the subject characteristics of documents;
- b) translating these characteristics into indexing terms;
- c) recording the terms with citation/ location; and
- d) arranging term in searchable order.

19.3 SEARCH STRATEGIES AND PRE-REQUISITES

Search strategy is the action plan which is drawn to conduct a search. It encompasses several steps and levels of work in information retrieval. There are many issues that need to be considered while formulating an appropriate search statement. These are:

- i) the concepts or facets to be searched and their order;
- ii) the term(s) that appropriately represent(s) the search concept;
- iii) the feature(s) of the retrieval system concerned; and
- iv) the measures to be taken in revising a search statement.

Developing a good search strategy requires knowledge about the nature and organization of target database(s) and also the exact needs of the user. Knowledge of the user's exact requirement can greatly affect the actual search and retrieval process. In some cases, the user may want only a few relevant items on a given topic, in which case the task of searching will obviously be limited. Conversely, the user may wish to obtain all the relevant items (obviously with as small a number of non-relevant items as possible), in which case the search must be exhaustive. Thus, an information search may fall in one of the following three categories [Chowdhury, 2004]:

- a) *High recall search*: when the user needs to find out all the relevant items on the stated topic. Recall is a parameter used to measure the performance of information retrieval systems; it is measured as the proportion of relevant items retrieved from a collection in a given search session;
- b) *High precision search*: when the user needs only relevant items, i.e., as small a number of non-relevant items as possible. Precision is a parameter used to measure the performance of information retrieval systems; it is measured as the proportion of the retrieved items that are relevant in a given search session; and
- c) *Brief search*: when the user wants only a few relevant items as opposed to all the relevant items.

19.3.1 The Pre-Search Interview

The results of a search depend heavily upon the correct understanding of the users' precise needs. This understanding can be developed through a pre-search interview. A pre-search interview is a conversation that takes place between a user and member of the information staff regarding the actual information requirement of the user. In order to conduct a successful pre-search interview, the interviewer should possess the following qualities:

- 1 ability to conduct personal communication;
- 1 conceptual skills;
- 1 analytical skills;
- 1 knowledge of file organisation;
- 1 understanding of indexing policy and vocabulary control; and
- 1 subject knowledge.

It may be noted that the concept of pre-search interview presupposes the existence of a search intermediary. This concept was developed to get a better understanding of the search requirements of a user in an online search environment. The concept of pre-search interview, although is very important in a mediated search process, has very little relevance in the context of information retrieval from the World Wide Web and digital libraries, since these systems are designed to be used by the end users without any direct involvement of human intermediaries. Nevertheless, an understanding of the activities involved in a search interview process may be quite useful for designers of information retrieval systems in the web and digital library environment.

The involvement of intermediaries reduces the uncertainty of retrieval to a great extent, thus, providing effectiveness in retrieval. The functions of the intermediaries are to ascertain exact information needs of the users as far as possible, identifying the concepts that describe the needs, translating these concepts into the terms of IR system and delineating the relationships among them, stipulating the parameters of the search. This, they do by interpersonal communication and participative discussion. The other functions are selection of database(s) for search, identification of alternative sources, formulation of alternative search strategies, matching of formulated search expression with the features of database(s) and conducting past search review. Many times, these are done by users interface. According to Taylor [1968], queries of a searcher shift from actual need to unexpressed need. The need is refined to the conscious description need or conscious need which is finally formalized as formalized need. The searcher's actual query formulation is done with interactive interface, in case of online systems and reference interviews.

Self Check Exercises

- 1) What is the difference between a high recall and a high precision search?
- 2) What is a pre-search interview in the context of online information retrieval?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

19.4 SEARCH TECHNIQUES

Different techniques are available to conduct searches effectively so that maximum relevant information are retrieved according to users need. Such techniques are:

19.4.1 Keyword and Phrase Search

A search can be conducted by entering a single search term or a phrase comprising more than one term. Keyword search is the simplest form of search facility offered by a search system. In the keyword search mode, the system searches the inverted file (the index) for each keyword/term forming the search expression. The search terms can be entered through the keyboard or can be selected from an index or vocabulary control tool, such as subject headings lists or thesauri. More than one keywords, forming a search expression, have to be combined using the Boolean or proximity operators (discussed below).

In a phrase search, the system searches for the entire phrase rather than each individual keyword forming the phrase. Phrase searches can be conducted only on those fields that are phrase indexed. If the index file comprises only single terms, then phrase search cannot be conducted, except of course by using the proximity operators (discussed later in this Unit) where the system searches for each constituent keyword in the search expression separately, and retrieves only those records where the keywords occur consecutively (as prescribed by the proximity operator) like in a phrase. Usually only a few fields in a database are phrase indexed, and phrase search can be conducted only on those fields. A search phrase can simply be entered through the keyboard, or through the index file or vocabulary control tools like subject headings lists and thesauri.

Different search system provides different facilities for conducting keyword and phrases searches. For example, in DIALOG search one can simply enter a keyword or a phrase preceded by the search command (s or select or find) and can click on the search button to conduct the search. User can also restrict the search to one or more fields (for further details on DIALOG search). Some search engines, for example, in *AltaVista* a phrase search is conducted when a search expression is entered within double quotes.

19.4.2 Keyword and Subject Search

Many bibliographic information retrieval systems provide two types of search facilities for conducting an unknown item search: keyword search and subject search. A **keyword search** allows users to enter one or more keywords pertaining to their query. These keywords can be chosen by the user in any combination depending on the requirements and there are several search operators that can be used to combine several keywords to formulate a search expression. The search keywords can appear anywhere, or in one or more chosen fields, in the database records. A **subject search** is the one that allows user to submit a subject expression that reflects his/her information requirement. Such a search is conducted on the subject field that contains one or more assigned subject headings entered by the indexer while creating the record in the database. Thus, a record will be retrieved only when user's subject search expression exactly matches the subject heading assigned by the indexer. In order to standardize the process, and also to help the user identify the appropriate subject heading, information retrieval systems use certain tools, called vocabulary control tools.

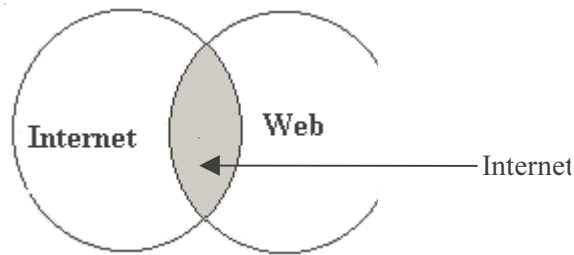
19.4.3 Boolean Search

This is a very common search technique that combines search terms according to the Boolean logic. Three types of Boolean search are possible: AND search, OR search and NOT search.

Boolean AND search allows users to combine two or more search terms using the Boolean AND operator. A Boolean AND search will retrieve all those items where all the constituent terms occur. For example, the following search expression "Internet and WWW" will retrieve all those records where both the terms occur. Boolean AND search adds more restrictions to a search expression by adding more search terms. Therefore, the more search terms are ANDed, the more restricted, or specific, will be the search, and as a result the less will be the search output. Sometime, a search may not produce any result if too many search terms are ANDed.

Example

Query: I am interested in the relationship between Internet and Web.

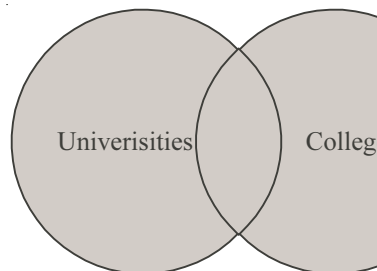


Search Items	Results	Remark
Internet	783	
Web	2962	
Internet AND Web	97	Shaded Area in the Figure

Boolean OR search allows users to combine two or more search terms such that the system retrieves all those items that contain either one or all of the constituent terms. Thus, the following search expression “Colleges or Universities” will retrieve all those records (1) where the term Colleges occurs, (2) where the term Universities occurs, and (3) where both the terms occur. Note that this is contrary to the use of the term ‘or’ in normal English. Boolean OR search, though adds more terms to a search expression, adds less restrictions to a given search expression, because the search is conducted for occurrence of each single ORed term irrespective of whether the other term(s) occurs or not. Consequently, the output of OR searches will be more. When too many search terms are ORed, the search output may be too big to handle.

Example

Query: I would like information about colleges and universities.

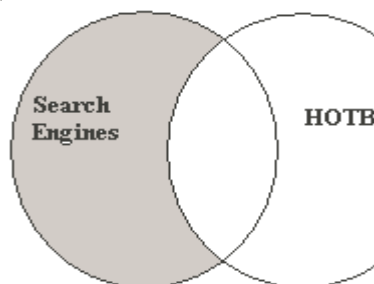


Search Items	Results	Remarks
Colleges	17320	
Universities	33685	
Colleges OR Universities	35702	Shaded Area in the Figure

Boolean NOT search allows users to specify those terms that they do not want to occur in the retrieved records. For example, the following search expression “Search engines NOT Hotbot” will retrieve all the records on search engines except those where the term ‘Hotbot’ occurs. Boolean NOT searches add restrictions to a search by forcing the search system to discard those items where the NOT term(s) occur. Hence the search output will decrease with increase in the NOT terms. The operators used for conducting a Boolean search vary from one search system to the other.

Example

Query: I want information about Search engines, but I want to avoid anything about Hotbot.



Search Terms	Results	Remarks
Search Engines	3651	
Hotbot	4556	
Search Engines NOT Hotbot	3001	Shaded Area in the Figure

In some search engines there are two approaches to conduct a Boolean search. The usual approach is by combining the search terms using the Boolean AND, OR, NOT, or any corresponding operators. For example, in *AltaVista* and *HotBot*, Boolean searches can be conducted using Boolean AND (or &), OR (or |) and NOT (or !), and combining terms and operators using parentheses. However, most search engines have another way to conduct search that implies Boolean search. For example, the plus operator (+) placed before a word or phrase means that all returned pages should contain that search term, and thus if the '+' symbol is used before two search terms, then the result will be as good as a Boolean AND search. Similarly the minus operator (-) can be placed before a word or phrase to exclude all documents containing that search term, and this implies the Boolean NOT search.

19.4.4 Truncation Search

Truncation is a search facility whereby a search can be conducted for all the different forms of a word having the same common root. As an example, the truncated word **COMPUT*** will retrieve items on **COMPUTER**, **COMPUTING**, **COMPUTATION**, **COMPUTE**, etc. A number of different options are available for truncation, viz., right truncation, left truncation, and masking of letters in the middle of the word. Left truncation retrieves all words having the same characters at the right hand part, e.g., ***hyl** will retrieve words like **'methyl'**, **'ethyl'**, etc. Similarly, middle truncation retrieves all words having the same characters at the left and right hand part. For example, a middle truncated search term **'col*r'** will retrieve both the terms **'colour'** and **'color'**. A 'wild card' is used to allow any letter to appear in a specific location within a word.

Right truncation and character masking or wild card are the most common truncation search facilities available in search systems. Operators used for truncation search vary from one information retrieval system to another; the most commonly used truncation operators include: *, \$, !, and ?.

19.4.5 Proximity Search

This search facility allows user to specify (1) whether two search terms should occur adjacent to each other, (2) whether one or more words occur in between the search terms, (3) whether the search terms should occur in the same paragraph irrespective of the intervening words, and so on. The operators used for proximity search and their meaning differ from one search system to the other. A proximity

search is as good as a Boolean And search in the sense that it searches for the occurrence of two or more search terms in the documents. However, it adds more constraints by specifying the distance between the search terms, and therefore the search output becomes more specific.

Example

```
sun within 4 words after moon
sun within 3 words before moon
sun within 2 words of moon
ice within 2 words after fire
ice within 4 words before fire
fire within 5 words of ice
```

19.4.6 Field-Specific Search

A search can be conducted on all the fields in a database, or it may be restricted to one or more chosen fields to produce more specific results. Specific fields and codes vary according to the search systems and database. The following examples show some valid DIALOG searches that have been restricted to some specific fields. The general format for using suffix codes is “**Syntax:** SELECT <term>/xx,xx ... where xx is a Basic Index field code(s)”

Select computer?/TI	Terms searched in the Title (/TI) field only.
S (information OR Communication)/DE,ID	Terms searched in either the Descriptor (/DE) or Identifier(/ID) field.
S S12/TI,AB	Restricts set S12 to either the Title (/TI) or Abstract (/AB) field.

In some cases one can use some prefix codes to restrict a search in a specific field. For example, in DIALOG one can enter the following search expressions to restrict the search in author or corporate source:

```
Select AU= Chowdhury,G
```

19.4.7 Limiting Search

Sometimes a user may want to limit a given search by using certain criteria, such as language, year of publication, type of information sources, and so on. These are called limiting searches. Parameters that can be used to limit a search are decided by the database concerned. The following are some examples of limiting searches in DIALOG:

<u>Limit Restriction</u>	<u>Qualifier</u>	<u>Example</u>
English-language documents	/ENG	SELECT URBAN(S)CRIME?/ENG
Patents	/PAT	S TRANSISTOR?/PAT

19.4.8 Range Search

Range search is very useful with numerical information. It is important in selecting records within certain data ranges. The following options are usually available for range searching, though the exact number of operators, their meaning etc., differ from one search system to another:

- 1 Greater than (>)
- 1 Less than (<)

- 1 Equal to (=)
- 1 Not equal to (≠ or <>)
- 1 Greater than or equal to (>=)
- 1 Less than or equal to (<=)

The following examples of range search are from DIALOG:

Publication Year	/yyyy	S Internet/2004
	/yyyy:yyyy	S Internet/2003:2004

Self Check Exercises

- 3) How does a truncated search improve the results of a search?
- 4) Which of the two search expressions will produce more hits (search results) and why?
 - a) Classification and indexing
 - b) Classification or indexing
- 5) What is the difference between a Boolean search and a proximity search?
- 6) What is the difference between a keyword search and a subject search?

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

.....

19.5 THE INFORMATION SEARCH PROCESS

19.5.1 Basic Features

Whether it is manual or online searching, basic features of search processes are same. These are:

- a) matching of information needs with the information available in IR system;
- b) conducting the search with some criteria such as surrogates, exact form, etc.; and
- c) subject surrogates are the main approach of information seekers.

19.5.2 Search Tactics

The success of any search depends upon the tactics effectively adopted in conducting a search. Search tactics is concerned with adopting techniques in maneuvering the actual search. The tactics described by Bates can be grouped in four categories [Bates, 1979]:

- a) Monitoring tactics (to keep the search on a track)

- b) File structure tactics (for traversing information within the system)
- c) Search formulation tactics (to aid in the process of designing and redesigning the search formulation)
- d) Term tactics (to select and reuse search terms)

In order to conduct a successful online search one must perform the following tasks:

- a) Decide which particular database(s) is/are to be searched;
- b) Guess the words that might have been used by the authors and the indexers in a database of potentially relevant documents;
- c) Use the thesaurus of the chosen database in order to translate the query terms in the appropriate way (is in the language of the system);
- d) Coordinate the terms (often using Boolean operators) to formulate the search statement;
- e) Use the search features and search operators appropriate for the chosen database;
- f) Input the search statement;
- g) Repeat steps (e) and (f) until a desirable output is obtained or the search fails altogether; and
- h) Identify the actual relevant items from among those retrieved.

One major task in the searching process relates to the coordination of terms [step (e) above] in order to formulate the actual search statement. The result of the search depends largely on how adequately the search terms are combined. Boolean search techniques have been used widely since the beginning of mechanised information retrieval. However, other models, especially the vector-space model (discussed in Unit 18 of this Course), or a variation of it is also used in modern information retrieval systems. In any case, the basic information search process that a user has to follow remains the same for almost all online information retrieval systems. In the following sections we shall see the major steps to be followed while conducting an information search using one of the oldest and most widely used online search services, viz. the Dialog information services.

19.6 ONLINE SEARCHING

Online information retrieval involves searching remotely located databases through interactive communication with the help of computers and communication channels [Chowdhury, 2004]. The database can be accessed by the user directly or through a vendor (supplier of online services); in each case through the computer and communication network. The term online retrieval can thus be used to indicate the information retrieval services available from producers of databases, or vendors of these databases. Although online information retrieval systems have existed for more than four decades, emergence of the Internet and world wide web have brought significant changes and improvements in the online information retrieval environment.

The phrase online searching was originally used to describe the process of directly interrogating computer systems to resolve particular requests for information. Now the phrase is used to denote searches that are conducted by means of a local computer that communicates with a remote computer system containing databases. Users can access the database(s) via an online search service provider (also called vendor). The search process is interactive and the user can conduct the search iteratively until a satisfactory result is obtained.

With the advent of the Internet and world wide web, the connotation of online searching has changed. Now we can conduct online searches through the world wide web on information sources that are distributed all over the world. For searching these information sources through the web, we can go straight to the web page of the service provider, provided we know the URL (uniform resource locator, or the address of the web page). Alternatively, we can try to locate the information source(s) by searching through the web search engines (the retrieval programs that help us search the web) like AltaVista and InfoSeek, or through subject directories or gateways (these are hierarchically organized lists that can be navigated to reach a particular information source or a group of similar sources) like Yahoo, SOSIG and Biz/ed.

In the following sections we shall first discuss the former type of online service – the traditional online service characterised by a remote online database search service offered commercially by a search service provider or vendor. We shall look at the online information search process with Dialog Information Service as an example. We shall then discuss the information retrieval from the world wide web with a view to understanding how traditional online searching differs from web information retrieval services.

One major advantage of the traditional online search services is that they are designed to be pay-as-you-go, and therefore each search session can be costed. Another advantage of online searching is its speed and the currency of the data retrieved. Originally online search services were very expensive and could be complex, and therefore intermediaries were needed to help end-users conduct an effective and efficient online search. However, over the years online search services have become less expensive and more user-friendly. As a result, they can now be used by end-users themselves, and can be accessed through the web.

19.6.1 Online Search Services

There are various components of an online search service such as:

- 1 **information providers** or database producers who provide databases to be accessed in an online mode;
- 1 **a search service provider** or vendor, which provides access to the databases and software for conducting the search;
- 1 **communication links** that connect the user with the host and the database(s); nowadays users can communicate with the service providers through the web and therefore an Internet connection is necessary; and
- 1 **a local workstation** through which the user is linked to the service.

Online search services, or vendors, are those organisations that provide value-added processing to the databases and offer search services. The following are some examples of online search services:

- 1 **Dialog** (<http://www.dialog.com/about/>): A pioneer in online search services, Dialog provides online access to over 800 million records in 900 databases in different disciplines.
- 1 **OCLC FirstSearch** (<http://www.oclc.org/firstsearch/>): This provides library users with instant online access to more than 72 databases, including these valuable OCLC databases: OCLC WorldCat, OCLC FirstSearch Electronic Collections Online, OCLC ArticleFirst, OCLC PAIS International, OCLC PapersFirst, OCLC ProceedingsFirst and OCLC Union Lists of Periodicals.

- 1 Ovid (<http://www.ovid.com/site/index.jsp>): Ovid provides access to hundreds of full text journals, renowned textbooks and premier bibliographic databases in various disciplines.
- 1 STN (<http://www.cas.org/stnonline.html>): STN offers current and archival information from over 200 scientific, technical, business and patent databases covering a broad range of scientific fields, including chemistry, engineering, life sciences, pharmaceutical sciences, biotechnology, regulatory compliance, patents and business.

While the above are examples of online search services that provide online access to a large number of databases in various disciplines, the following are examples of some online search services that provide access to the full texts of journals and books:

- 1 EBSCO Information service (<http://www.epnet.com/default.asp>): This provides access to a large collection of full text and bibliographic databases suitable for all kinds of libraries.
- 1 Ingenta (<http://www.ingenta.com/>): Ingenta provides access to the full text of over 5,400 publications from over 230 academic and professional publishers.
- 1 ProQuest (<http://proquest.umi.com/pqdweb>): ProQuest is a resource of electronic collections containing millions of articles originally published in magazines, newspapers and journals.

Details of online search services are available in a number of publications [see for example, Forrester and Rowlands, 1999; Large, Tedd and Hartley, 1999; and Chowdhury and Chowdhury, 2001b] as well as in the websites of the respective search services. Various features of online search, and the various steps involved in carrying out an online search are discussed below, with examples from DialogWeb.

Self Check Exercises

- 7) What is meant by online searching?
- 8) Give three examples of online search services.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

19.6.2 Basic Steps in an Online Search

The steps involved in carrying out an online search vary from system to system. This is because each system has its own custom-built interface, which allows specific types of search and uses specific operators for different search commands. Nevertheless, the graphical user interfaces used in these systems have made the task of searching reasonably straightforward and the process of searching has been simplified further in the web-based interfaces. These are the basic steps that one needs to follow to conduct an online search:

- i) *Study the search topic and develop a clear understanding of the information requirement.* This is a critical step and depends on a number of factors, such as,

the nature and requirements of the user, how well the user can express his or her information needs, how much the user already knows, how the user is going to use the information, and so on. This happens before the actual search process begins and is often conducted through a series of dialogues between a searcher and an information intermediary. In the absence of an intermediary, users have to clearly delineate their information requirements for themselves. This stage decides the strategies to be adopted for a search.

- ii) *Get access to an online search service.* This can be done through subscription or a licensing agreement. The access right has to be obtained before the search begins.
- iii) *Log on to the service provider.* Nowadays this is usually done through the web interfaces of the online search service providers. Users need to know the URL of the online service provider as well as the user login-ID and password. Sometimes Internet Protocol (IP) authentication mechanism is adapted instead of user login-ID and users directly logon to the portal.
- iv) *Select the appropriate database(s) to search.* This is a critical and often a difficult task. The success of a search largely depends on the appropriate selection of the databases. Online search services allow users to select one or more databases to search using the same interface. Most search services allow users to browse through the database categories to select appropriate databases(s). Dialog has a unique facility called DialIndex search (details are given later in this unit), which allows users to see how many times a given search term occurs in a set of chosen databases. This information can guide users to select the appropriate database to conduct the actual search.
- v) *Formulate search expressions.* This is the key part of the job. It may involve a number of activities, the first being the selection of appropriate terms and/or phrases. This may require the user to consult dictionaries and thesauri. Once the appropriate search terms and/or phrases are chosen, the search expression has to be formulated. At this stage the user should have an understanding of the nature, content and structure of the chosen database(s) and to know which fields are indexed and therefore can be searched. The user also needs to know what search facilities are available, such as Boolean search, truncation, field specific search, proximity search, and so on, and the appropriate operators. The search operators and syntax for formulating search expressions vary from one search service to the other. Many search service providers have different interfaces for novice and expert users. If the users want to use the expert search interface, which may be command-driven, they have to have a knowledge of the various search commands and their order of execution.
- vi) *Select the appropriate format for display.* Online search services allow users to select an appropriate format, from a number of predefined formats, to display the retrieved records. However, there may be charges for the records displayed. For example, when searching Dialog, charges incurred include output and search time costs, as well as Internet charges; prices also vary by database. Therefore, one has to be very careful in deciding which record(s) display and in which format to display them. If the option for the display of the full record(s) is chosen, the process may take some time, depending on the network traffic. However, each online search service provides an option for brief display, which shows the brief details of the output records, and users may select records from this list for a full display.
- vii) *Reformulate your query, if necessary.* This may mean going back to step 4 or step 5 and repeating the entire process. Online searches are usually iterative processes, meaning that user conducts several searches, compares the results, modifies a search statement, or conducts a new search in order to get the best results.

viii) *Select the mode of delivery.* You may download all the chosen records online or send an offline request.

In the following section, the methodology of DialogWeb search is provided to give you an idea about how online search can be conducted.

Self Check Exercise

9) What the major steps in an online information search?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

19.6.3 Features of an Online Search Service: DialogWeb

DialogWeb is the web interface to the Dialog online search service, one of the oldest and largest online search service providers, which gives easy access to a large number of databases with:

- 1 company information – both directory listings and financial information
- 1 industry information – trends; overviews; market research; specialised industry newsletters and reports; US and international news, including an extensive collection of newspapers and newswires from North America and Asia; and US government news, including public affairs, law and regulatory information
- 1 patents and trademarks – a worldwide collection for research and competitive intelligence tracking
- 1 chemistry, environment, science and technology – technical literature and reference material to support research needs
- 1 social science and humanities including education, information science, psychology, sociology and science, from public opinion, news, and leading scholarly and popular publications
- 1 general reference information – people, books, consumer news and travel.

Users can search and retrieve information from all these different types of information sources using:

- 1 guided search mode, which does not require knowledge of the Dialog command language;
- 1 command search mode, which allows experienced users to use the Dialog command language;
- 1 database selection tools, which help users pinpoint the right database for a search;
- 1 integrated database descriptions, pricing information and other search assistance; and
- 1 easy to use forms to create and modify Alerts (current awareness updates).

Dialog search results are available in HTML or text formats. Users have a choice of displaying records or sending search results via email, fax, or postal delivery.

19.6.3.1 Steps in a DialogWeb Search

The first step of a DialogWeb search involves logging in to the system, for which

a Dialog account is necessary. The user goes to the DialogWeb site (<http://www.DialogWeb.com>) and must enter the user ID and password. The log-in screen also provides information about DialogWeb and a preview and search tips. After logging in, the user needs to select the mode of search: Guided Search or Command Search. Guided Search is the default search option.

19.6.3.1.1 Guided Search

Guided Search is designed for novice to intermediate searchers. The following steps are to be followed for conducting a Guided Search looking for information on the topic of 'digital libraries'.

Step 1: Choose Database

To begin the Guided Search, the user clicks the New Search button and chooses from the list of main subject categories. Each category is further divided into focused search topics. For a search on digital libraries, one can select these categories:

Social Sciences and Humanities > Social Sciences > Library and Information Science

This will lead to a list of databases that cover Library and Information Science.

Step 2: Choose a Search Option and Carry Out Search

In Guided Search there are two search options:

- 1 *Targeted Search*, which is available in some, but not all, subject categories. It is a ready-made search form with databases pre-assigned to the form.
- 1 *Dynamic Search*, which is available in all the subject categories. The Dynamic Search form is generated based on the category or database that is selected. Dynamic Search has access to many more databases compared than the Targeted Search and is more flexible.

Targeted Search is the easiest type of search to perform. The user can enter the search word or phrase as 'Words in Title' or as the 'Main Subject'.

Dynamic Search is available at various points in the search category selection process or when a user chooses the Quick Functions option in New Search and enters a specific database number. The Dynamic Search capability is available no matter what category or database is picked. In a category with many databases assigned to it, a user can search:

- 1 all of the databases together
- 1 a group of similarly designed databases together
- 1 one of the assigned databases individually.

If a user has chosen the Dynamic Search option and has decided to conduct the search on all the databases under the 'Library and Information Science' category, the 'Dynamic Search' screen is shown. The Dynamic Search forms also offer the following options:

- 1 Navigation – The search category selections display at the top of the form. To return to a category or option, the user clicks the search category or option name.
- 1 Run Saved Strategy – If a user has already saved a search strategy, it can be run against the selected databases by clicking Run Saved Strategy.

A list of the databases used in the search is displayed at the bottom of the form. The info (i) icon gives more information about the database content and pricing.

In the Dynamic Search screen users can enter a search term or phrase and conduct the search on subject, author, descriptor or title field. A search can also be restricted by the year of publication, and the user can browse the list of items by author or year of publication.

Step 3: Display Search Results

The search results from a Targeted Search or a Dynamic Search will appear on a Picklist page, which provides a quick view of the records. From the Picklist page users can choose to:

- 1 display specific records in more detailed formats or send records via e-mail or fax, or by post
- 1 rearrange the order in which the records are displayed
- 1 refine the search strategy
- 1 remove duplicate records
- 1 view the prices for all format options
- 1 save the strategy for future use
- 1 create an Alert for automatic updates on the search topic.

After the search has finished processing, the Picklist page will appear. Users can choose to view results by selecting one or more items by checking the boxes and then selecting the display button, or can display any one record just by clicking on the hyperlinked title. The format for display is chosen from the 'Format' box and the records are sorted according to a sort criterion chosen from the 'Sort by' box. The search expression can be refined by clicking the 'Back to Search' button, which allows users to edit, add or delete information from the search form.

19.6.3.1.2 Command Search

Command Search is designed for intermediate to experienced Dialog searchers. It provides complete command-based access to Dialog's extensive collection of databases. Users are expected to be familiar with the various Dialog commands when using Command Search. Additional features include built-in tools such as Bluesheets (database descriptions) and pricing information, database selection assistance to help pinpoint the right databases for a search and easy to use forms to create and modify Alerts (current awareness updates). The Command Search main page allows users to begin inputting Dialog commands immediately. A Command Search contains:

- 1 a textbox for entering Dialog search commands
- 1 a Submit button that sends the command
- 1 a Previous button that displays your most recent command entries.

The main page has links to the Databases feature, product support information, and Guided Search. Users can move between Guided Search and Command Search. Steps for conducting a Command search can be summarised as follows.

Step 1: Choose Database(s)

DialogWeb simplifies database selection by arranging the databases by subject in the Databases feature. Users can select one or more databases by checking in the Database box. However, if users are not sure which database(s) to select, they can choose the Dialog Index option. This is particularly useful when users do not know which databases to search, or when they want to carry out a comprehensive search and cover everything on a topic. Dialog Index is a master index to most of the Dialog databases, and it allows users to compare the number of records retrieved from a group of databases.

After selecting a database, users must search the databases to view the records. They can click 'Begin Databases' to enter the files that they have checked and run the same strategy, or may choose the database(s) to search by entering the file numbers and even change their search strategy in the command line.

Step 2: Choose a Search Option and Carry Out Search

Once the databases are chosen, the Dialog Command Search page appears. It can also appear:

- 1 after log-in if it is set as the default
- 1 when the Command Search link from the main Guided Search page is clicked
- 1 when the Begin Databases button from Databases is clicked for browsing.

The appropriate BEGIN or 'b' command is inserted in the command line automatically when a search has been made in Databases and one or more databases have been selected. Users can add the CURRENT command to their BEGIN statement by typing in 'current' after the command. This allows them to search the current year and one year earlier, and narrows the search results at the beginning. Then they click the Submit button or press the ENTER key on the keyboard to start the search.

Any Dialog command followed by the search term(s) should be entered in the search box. The terms when looking for information on digital libraries might be: S digital (w) libraries.

Step 3: Add Operator to a Search

The search can be refined by including 'electronic libraries' through the following expression:

S (digital or electronic) (w) libraries

This search statement will retrieve records on electronic as well as digital libraries. The search statement retrieves those records where digital and libraries, or electronic and libraries, occur next to each other in the same sequence. More records are retrieved by truncating the search term libraries as follows:

S (digital or electronic) (w) library?

Various other modifications may be made by using appropriate search commands, for example, limiting the search to one or more fields or limiting the results to a language, year of publication, and so on.

Step 4: Displaying Records

A search history of all of the sets appears and users can view some of the records. It is a good idea to display a few records in 'free' format before displaying the records in the long or full format. To display records users can choose a format from the drop-down list and click Display for the appropriate set. Formats determine the amount of information to be displayed for each record. The Format list box lists the basic format options: free, short, medium, long, full and KWIC. It is possible to indicate the number of records to display; the default is 10 and a maximum of 99 records can be specified. There is an option of using a Type command (see Table 1 for details) to display records or From Each together with the Type command, in order to search more than one database.

Dialog Search Operators

Dialog offers a number of search features, such as Boolean search, proximity search, truncation, field-specific search, limiting search, and so on. The various search features of Dialog and the corresponding operators are shown in Table 19.1.

Table 19.1: Examples of Dialog search operators

Search operator (can be used lower or upper case)	Example	Description
Proximity operator (w)	S information (w) system	Requests adjacent terms in the defined order.
	S information (2w) system	You can specify the maximum number of intervening words. In the second example, there may be at the most two words in between information and systems, but information should appear first followed by systems with at the most two intervening words
Proximity operator (n)	S information (n) system	Requests adjacent terms in any order.
	S information (2n) system	You can specify the maximum number of intervening words. In the second example, there may be at the most two words in between information and systems, but any term may appear first
Proximity operator (s)	S solar (s) energy	Requests that terms should be in the same subfield unit
Proximity operator (l)	S solar (l) cell	Requests that terms should be in the same descriptor unit
Proximity operator (F)	S information (F) system	Requests that terms should be in the same field
Open truncation ?	S computer?	This is an open truncation; any number of characters can appear after the root word
Truncation operator ? ? (note that there is a space in between the two operators)	S cat? ?	A maximum of one character can appear after the root word
Truncation operator ?? (note that there is no space in between the operators)	S comput??	As many characters as the number of '?' (in this example two) can appear after the root word
Internal truncation ?	Wom?n	Will replace the '?' with one character
Field specific search		Suffix codes can only be used for the basic index fields
Suffix code	S internet/ti,de	Prefix codes are used to search the additional indexes
Prefix code	S au=Chowdhury, G.	
Limit suffixes	S internet/1999	Used with the select command to restrict the search by a given criteria
Range searching	S internet/2002:2003	The search can be restricted to a range of dates
Numeric searching	S EM=100	In the first example the system will retrieve records with 100 employees; in the second with employees of more than 1000
	S EM>1000	
Boolean AND	Information AND Internet	Retrieves those records where both the search terms occur
Boolean OR	Information OR Internet	Retrieves those records where either or both the search terms occur
Boolean NOT	Internet NOT web	Retrieves those records where the first search term occurs but not the second

Dialog is a pioneering online information service and now many new online search services have better user interfaces and search and retrieval features.

Self Check Exercises

- 10) What is the basic difference between a guided search and a command search in Dialog?
- 11) What is difference between the following two search expressions in Dialog:
- S (digital or electronic) (w) libraries
- S (digital or electronic) and libraries

Note: i) Write your answers in the space given below.
 ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

19.7 ACCESS TO INFORMATION ON THE WEB: THE TOOLS

A user can get access to any website by entering the URL (Uniform Resource Locator; the address of a web site) on the browser. A web browser, like Netscape Navigator or Microsoft Internet Explorer, is a computer program, an essential tool for getting access to the web. A web browser performs two major tasks:

- 1 It knows how to go to a web server on the Internet and request a page so that the browser can pull the page through the network and into your machine.
- 1 It knows how to interpret the set of HTML (Hypertext Markup Language; a language format used to create web pages; HTML is called the *lingua franca* of the web) tags within the page in order to display the page on your screen as the page's creator intended it to be viewed.

Although one could get access to any web page by typing on the browser the URL of a sought website, and then moving into the site through the various links, there are several problems to this approach, especially when the user is interested to get some specific information on a given topic, or find answers to a given question. Problems arise because it is almost impossible for users to know which of the billions of web pages may contain the information they require and which of the millions of websites contains the required web page. In order to solve these difficulties, several web search tools have been developed that assist users in finding the information they need from the right web page with relatively little effort.

There are basically two ways to find information on the web: by conducting a search using what is known as a search engine, or by following the links in a specially designed list called a directory. Search engines allow users to enter search terms – keywords and/or phrases – that are run against a database containing information on web pages collected automatically by programs called spiders. The search engine retrieves from its database web pages that match the search terms entered by the searcher. It is important to note that when a user conducts a search using a search engine, the search engine does not search for the information across the entire web at the given instance. Instead, it searches a fixed database, which is updated at a regular intervals according to a specific set of criteria employed by the search engine, located at the search engine's website and containing information on selected web pages.

19.8 HOW THE SEARCH ENGINES WORK

Although all search engines are intended to perform the same task, each goes about doing so in a different way, sometimes with very different results. Factors that influence the search results include the size of the database, frequency of updating it, criteria employed for indexing the chosen web pages, and the search engine's retrieval capabilities. Search engines also differ in their search speed, the design of the search interface, the way in which they display the search results, the amount of help they provide, and in other ways.

Search engines run from special sites on the web and are designed to help people find information stored on other sites. There are differences in the ways search engines work, but they all perform the following three basic tasks:

- 1 They search the Internet – or select parts of the Internet – according to a set of criteria.
- 1 They keep an index of the words or phrases they find, with specific information such as where they found them, how many times they found them, and so on.
- 1 They allow users to search for words or phrases or combinations of words or phrases found in that index.

There are three main components of a search engine: the spider, the search engine software and interface, and the index.

19.8.1 The Spider

To find information from the millions of web pages, a search engine employs special software called a spider or crawler. It is a program that automatically fetches web pages for search engines; it is called a spider because it *crawls* over the web. Spider programs treat the web as a graph and, using a set of URLs as a seed set, traverse the graph to select web pages. The crawler traverses the graph either breadth first (searching all nodes at one level of the tree before going down a level) or depth first (searching the current path as far as possible before backtracking to the last choice point and trying the next alternative path in the tree). Web pages contain links to other pages, and a spider uses these links to move to another page; it visits the page, reads it and then follows links to other pages within the site.

One of the major problems for a spider or a crawler program, and indeed for a specific search engine, is to decide which page to select for indexing. Each search engine aims to index the most important web pages, and therefore aims to prioritize URLs to obtain the best pages. The quality of a web page may be judged in many ways, for instance by measuring its content, by assessing its popularity (by counting the number of visits) or by measuring its connectivity (which other pages link to this page). Spider program metrics based on connectivity have the advantage that they do not require information that is not easily accessible (such as page popularity data), and that they are easy to compute, so they scale well to even very large page collections. Another major issue for a crawler program is to schedule the frequency of revisiting pages. Since web pages keep changing constantly it is important to visit them frequently. Search engines do not usually disclose the details of their spider programs.

19.8.2 Search Engine Software

Search engine software is the information retrieval program that performs two major tasks: it searches through the millions of terms recorded in the index to find matches to a search and it ranks the retrieved records (web pages) in the

order it believes is the most relevant. The criteria for selection (or rejection) of search terms and assigning weight to them depend on the policy of the search engine. Similarly, the specific information that is stored along with each keyword – such as where in a given web page it occurred (heading, links, meta-tags or title of the page), how many times it occurred, the attached weight, and so on – depend on the policy of the search engine concerned [Sullivan, 2003]. Each commercial search engine has a different formula for assigning weight to the words in its index. This is one of the reasons for the fact that a search for the same word on different search engines may produce different results, and the retrieved web pages may be ordered differently.

Google uses the concept ‘page rank’ to determine the importance of a web page [Brin and Page]. The idea is based on the principle of citation analysis. A document, according to the basic principle of citation analysis, is considered to be important if it is cited frequently, and thus one can rank a set of documents in order of their importance by counting the number of times each one has been cited. A web page’s ‘page rank’ is an objective measure of its citation importance that corresponds well with people’s subjective idea of importance.

19.8.3 Indexing

Early search engines indexed only components of each web page, but increasingly full texts of web pages are indexed [Rasmussen, 2002]. Specific information on the form and weight of index terms, the techniques for calculating relevance, and so on are usually proprietary. Most search engines use variations of the Boolean and vector space model. Although search engines do not usually disclose their secret of ranking pages, some general information is available. One of the main rules in a ranking algorithm involves the location and frequency of keywords on a web page [Sullivan, 2003]. The location of a term on a page is also used as an important criterion. Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant to the topic than others. Pages where the search terms appear near the top of a web page, such as in the headline or in the first few paragraphs of text, are also ranked highly compared with others. Frequency of occurrence is the other major factor used to determine relevance. Pages where the search terms occur frequently are often deemed more relevant than other web pages.

Features of each search engine can be learnt by following the help pages. However, there are a number of sources that regularly report on the features and comparisons of web search engines. The most prominent are *SearchEngineWatch.com* and the *Online* journal.

19.9 COMMON SEARCH AND RETRIEVAL FEATURES OF WEB SEARCH ENGINES

Almost all search engines provide basic text search facilities, like, Boolean search, proximity search, phrase search, truncation, field-specific (often called meta tag) search and limiting search (by language, domain, country, and so on), as shown in Table 19.2. The search operators vary from one search engine to another. However, the search engine interfaces have improved significantly in ease of use compared with the interfaces of many traditional text retrieval systems.

Table 19.2: General Search Features of Search Engines

Search features	Examples
Boolean search	Usually three different approaches are followed: <ol style="list-style-type: none"> 1 combining search terms with AND, OR, NOT (e.g., ANDNOT in AltaVista) 1 using '+' or '-' signs before the search terms 1 selecting options like 'all of the words' (in Google); or 'any of these words' or 'none of these words' (in AltaVista)
Proximity search	Using an appropriate operator, e.g., 'NEAR' (in AltaVista)
Field search, e.g., title search	Using 'intitle:' or 'all in title:' before the search terms in Google
Phrase search	Entering a phrase within double quotes (available in AltaVista, Google), or by selecting an appropriate option (e.g., 'with all of the words' (in Google), 'this exact phrase' in AltaVista)
Limiting search	Limiting by time, date or file type (in AltaVista); by date, file format, language or occurrences (in Google)

It may be noted that one of the major improvements in the web search environment has been in the design of user interfaces that allow users to conduct complex search queries without having to remember or use complex search syntax or operators. Many advanced search screens also provide enough help information for users to complete their search on the search screen itself.

In addition to those retrieval features that are common to the traditional text retrieval systems, many search engines provide some unique search facilities that have been designed specifically for web information retrieval. Table 19.3 shows some unique features of search engines and provides examples of search engines where the feature is available.

Table 19.3: Special Features of Search Engines

Special Features	Examples
Host/domain search	In the advanced search screen of AltaVista users can type in a domain name from a list of domain and country codes; in Google, users can type in a domain name (URL) in the search box and select to get or not to get results from the domain
URL search: users can search within a site by typing in the top-level URL	In the advanced search screen of AltaVista and Google users can type in a URL in the search box
Link search: a page-specific search whereby users can specify a page and ask the search engine to retrieve those pages	In the advanced search screen of Google users can type in a that link to the given pageURL to find pages that link to the given page
Limiting search: users can restrict the search results by specifying a number of limiting factors	In AltaVista users can limit searches by date, file type and location; in Google a search can be restricted by language, file format, date, occurrences and domain
Find similar: this is a page-specific search whereby the user can specify a web page and ask the search engine to find similar pages; this option is also available in the search output pages of some search engines	In Google, users can enter the URL of a page in the advanced search screen. The 'Similar Pages' option in Google and 'Related Pages' option in AltaVista find pages similar or related to the one selected

Special Features	Examples
Stemming	In the advanced search screen of MSN, users can select the 'Enable Stemming' option to search for the variant forms of a word
Search by language	In the advanced search screen of Google and HotBot, users can choose a language from a list of languages
Filter: users can choose to activate a filter in order to stop the search engine from retrieving unwanted materials	In the advanced search screen of HotBot users can choose options to block offensive contents; in AltaVista and Google there is an option to activate the filter option in the settings
Page content: user can specify the content of a page	In HotBot users can specify the content type of the retrieved pages by selecting options like audio, image, MP3, Word, PDF, audio or video; in the advanced search screen of AltaVista, users can choose a file type: all, HTML or PDF file formats; in MSN users can select the file type or file content type from a list
Page translation	AltaVista and Google options (called 'Translate' in AltaVista and 'Language Tools' in Google) that allow users to enter text in one language and translate it into another language
Sorted by option (in AltaVista)	AltaVista allows users to enter search terms in an appropriate box to tell the search engine to sort the results based on the occurrence of those terms
Results display: users can select the number of output to be displayed on each page	In AltaVista and Google users can select the number from a list

In addition to those mentioned in Table 19.3, some search engines provide further special features to help the searcher. For example:

- 1 The 'I'm Feeling Lucky™' button in Google takes users directly to the first web page Google returned for a query.
- 1 Site Collapse in AltaVista shows a maximum of two results per site, allowing the user to look at results from various sites.
- 1 Toolbar (<http://toolbar.google.com>) is a special toolbar for Internet Explorer users that puts a Google search box right into the user's browser.
- 1 TouchGraph GoogleBrowser: Shows users what the web 'looks like' to the search engine, displaying the links between the user's favourite websites.
- 1 Yahoo Companion: Allows users to search Yahoo, access Yahoo Mail, checks on stocks and more via a toolbar within the browser.

19.9.1 Specialised Search Engines

In addition to retrieving text, search engines also provide facilities for retrieving multi media information. For example, the 'Images' option from the main webpage of Google takes the users to the Google image search page where the user can enter one or more terms for searching images. Similar facilities are also available from other search engines. For example, Altavista provides different search pages from image, audio and video search.

A number of specialized search engines provide alternative approaches to query formulation and the display of results. The two most useful options are the natural language query formulation facilities, and visual display of the search terms and/or search results.

19.9.1.1 Natural Language Queries

Some search engines allow users to enter a query in simple natural language. The idea behind natural language queries is that users can type a question in the same way that they would ask it of a human and therefore they don't need to keep track of Boolean or other search operators or complex query structures.

One of the most popular natural language query sites today is AskJeeves.com (<http://www.askjeeves.com>). Users can ask a simple natural language question and AskJeeves comes up with answers and/or a set of related questions, and/or hits from other search tools like AltaVista or Yahoo. One of the unique features of AskJeeves is that it takes users to the page containing the answer to their query. On top of the page an AskJeeves frame appears, which allows users to print the answer, modify the query, and so on. Another unique feature of AskJeeves is that, in addition to the results from various search engines, it comes up with a number of predefined queries related to the question that has been asked by a searcher. The searcher can select any one of them, click on the Ask button, and return the query to AskJeeves. This allows a searcher to pick up the most appropriate query with which to pursue their search. AskJeeves also has a list of popular questions and users can choose any one of them to return it to AskJeeves as a query. AskJeeves processes and categorizes all the queries that it receives, and presents all the related questions on to the screen. In most cases, unlike other search or meta search engines, AskJeeves takes the user directly to the website that contains the answer to their question.

19.9.1.2 Information Visualisation

A number of search engines allow users to view a visual display of the links among the retrieved web pages on the screen. The TouchGraph GoogleBrowser allows users to view a visual display of links between a chosen website and other sites. Users need to install a program (available free from the Google site) to activate this feature.

Many search engines allow users to enter a search query and produces a visual display of the retrieved output, as opposed to a text page listing the retrieved sites. A typical example of such a search engine is Kartoo (<http://www.kartoo.com>). The search results screen in Kartoo has two parts: on the right-hand side of the screen users can see a display of retrieved sites, under specific topic headings, linked to each other. By moving the mouse to any one node (website) on the screen, the user can see the link of the chosen site with other sites on the screen. On the left-hand side of the screen there is a list of the top sites and a list of topics that are related to the search term entered. Users can choose any of the listed topics to modify the search.

Another interesting search engine is Vivisimo (www.vivisimo.com), which shows search results clustered under different topics related to the search term(s). The search output screen has two parts: on the left results are displayed under various related topics, and on the right there is a set of pages retrieved by the search term(s). Wisenut (<http://www.wisenut.com>) also shows results retrieved

by the search term(s) together with a cluster of web pages on related topics.

Self Check Exercises

- 12) What is a spider in the context of a web search engine?
- 13) What is a meta search engine? Give example of two meta search engines.
- 14) Give examples of two web search engines that use visualization techniques for display of information on the screen

Note: i) Write your answers in the space given below.
 ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

19.10 MULTIPLE DATABASE SEARCHING AND COMMON COMMAND LANGUAGE

Databases are generated by various agencies in different subject fields and database producers use different commands and instructions for searching the databases. Different software packages also provide different command languages. As a result users need to be conversant with different command languages and have to have sufficient knowledge if they intend to search more than one database. As it is known, thousands of databases are now existing or are being developed. Thus, it becomes difficult for any user to develop mastery in all of them, even if he has to search small number of databases in his own subject/ discipline. Some online database vendors like Dialog, STN, who provide access to several databases, have solved the problem to some extent by adopting some command language to search the databases serviced by them.

To solve the problem of heterogeneity in command languages, some efforts are being made at international level. Z39G Committee of National Information Standards Organisation (NISO), USA, set up in 1980 came up with a draft Common Command Language (CCL) in 1987. The International Organization for Standardization (ISO) also came up with a standard CCL in 1988 (ISO 8777). There is great similarities between these two standards. However, implementation of these is still awaited. Wide implementation of the CCL will facilitate the users in reaching multiple databases of different producers and vendors with less effort. With this, for database producers and users it will be advantageous to pay more attention to the contents part while searching. It will also facilitate the trainers to pay more attention to the intellectual part of content of databases rather than spending more time on the technicalities of searching the databases.

19.11 SUMMARY

An information search is a complex process and it involves several stages and several decisions. Traditional online information services were designed as intermediary-based services, in the sense that they were expected to be used by the end users with the help of expert search intermediaries.

An online search involves several steps. Pre-search interview is a process that

helps the intermediaries understand the user queries in a better way. Dialog information services has been the pioneering and most widely used information search services. The steps involved in a Dialog search are quite distinct, and one needs to precisely follow those steps in order to get better search results.

The connotation of online searching has changed with the advent of the Internet and world wide web. Users now can get access to remote online information resources through the web. Several different types of web search tools have been developed over the past few years. Although these tools are free at the point of use, they have been constantly changing and improving their features. Characteristics of some general and special web search tools have been discussed in this Unit. Overall this Unit will give a clear understanding of the complexities and techniques of retrieval of information from remote online sources. Different database producers follow different command languages to search their databases. Development of Common Command Language will facilitate all concerned in the chain.

19.12 ANSWERS TO SELF CHECK EXERCISES

- 1) A high recall search aims to produce the maximum number of relevant items from a collection (even if in that process some non-relevant items are retrieved) while a high precision search aims to produce precisely the relevant items (the objective being to retrieve as many relevant items as possible, and at the same time filtering out non-relevant items).
- 2) A pre-search interview is a conversation that takes place between a user and member of the information staff regarding the actual information requirements of the user.
- 3) A truncation search retrieves all those records where part of the search term occurs. For example a truncation search with 'librar\$' will retrieve records with terms like libraries, library, librarian, etc. Thus, a truncation search produces more search results.
- 4) The second search, i.e., 'classification or indexing' will produce more results since this search will retrieve all those records where either or both the search terms occur, whereas the first search expression 'classification and indexing' will retrieve only those items where both the search terms occur.
- 5) A proximity search is as good as a Boolean and search in the sense that it searches for the occurrence of two or more search terms in the documents. However, it adds more constraints by specifying the distance between the search terms, and therefore the search output becomes more specific.
- 6) A **keyword search** allows users to enter one or more keywords pertaining to their query. The search keywords can appear anywhere, or in one or more chosen fields, in the database records. A **subject search** is the one that allows user to submit a subject expression that reflects his/her information requirement. Such a search is conducted on the subject field that contains one or more assigned subject headings entered by the indexer while creating the record in the database. Thus, a record will be retrieved only when user's subject search expression exactly matches the subject heading assigned by the indexer.
- 7) The phrase online searching was originally used to describe the process of directly interrogating computer systems to resolve particular requests for information . Now the phrase is used to denote searches that are conducted by means of a local computer that communicates with a remote computer system containing databases.
- 8) Dialog, Ovid, STN

- 9) The following are the major steps in an online search:
 - i) Study the search topic and develop a clear understanding of the information requirement.
 - ii) Get access to an online search service.
 - iii) Log on to the service provider.
 - iv) Select the appropriate database(s) to search.
 - v) Formulate search expressions.
 - vi) Select the appropriate format for display.
 - vii) Reformulate your query, if necessary.
 - viii) Select the mode of delivery.
- 10) A guided search is meant for novice users whereas a command search is meant for experienced or expert users. A guided search takes the user through the various stages of a Dialog search, whereas a command search expects the user to know the specific search commands and use those commands to conduct a search. A guided search is easy to use it often takes more time, and does not allow the users to create complex search expressions. However, a command search, although expects the user to know the Dialog commands, is faster and allows the users to create complex search expressions.
- 11) Both search statement will retrieve records on digital libraries or electronic libraries. In the second search, the two terms digital and libraries, or electronic and libraries can occur anywhere in the retrieved documents. Because of the proximity operator w, that specifies that the two terms should occur next to each other, the first search expression will retrieve only those records containing the phrase 'digital libraries' or 'electronic libraries', i.e., the terms digital and libraries should occur next to each other as it appears in a phrase, and the same thing will happen for the terms electronic and libraries.
- 12) To find information on the millions of web pages, a search engine employs special software called a spider or crawler. It is a program that automatically fetches web pages for search engines; it is called a spider because it *crawls* over the web.
- 13) Tools that allow users to conduct concurrent searches on more than one search engine. Examples: Dogpile, Mamma
- 14) Kartoo, Vivisimo

19.13 KEYWORDS

Meta Search Engine	: A search engine that allows users to run concurrent searches on more than one search engine.
Online Search Services	: Information providers or database producers who provide databases to be accessed in an online mode.
Spider or Crawler	: It is a program that automatically fetches web pages for search engines; it is called a spider because it <i>crawls</i> over the web.
Web Browser	: A computer program that is used to get access to the Internet, typical examples being Netscape Navigator and Microsoft Internet Explorer.
Web Search Engine	: Programs that allow users to search and retrieve information from the web.

: Wildcard Character, in computer science, a keyboard character that can be used to represent one or many characters; usually encountered with operating systems as a means of specifying more than one file by name. Example: the question mark (?), the asterisk (*), etc.

19.14 REFERENCES AND FURTHER READING

Bates, H.J. (1979). Information search tactics. *Journal of the American Society for Information Science*, 30, 205-47.

Brin, S. and Page, L. *The anatomy of a large-scale hypertextual web search engine*. <<http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>>

Chowdhury, G.G. (2004). *Introduction to modern information retrieval*. 2nd ed. London: Facet Publishing.

Chowdhury, G.G. and Chowdhury, S. (2001a). *Information sources and searching on the World Wide Web*. London: Library Association Publishing.

Chowdhury, G.G. and Chowdhury, S. (2001b). *Searching CD-ROM and online information sources*. London: Library Association Publishing.

Forrester, W.H. and Rowlands, J.L. (1999). *The online searcher's companion*. London: Library Association Publishing.

Ghosh, S.B. and Sarkhel, J.N. (eds.) (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Gopinath, M.A. (1999). *Search strategies and heuristics*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *The process of searching*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.

ISO 2788:1986. *Guidelines for the establishment and development of monolingual thesauri*. Geneva: International Organization for Standardization.

Large, A.; Tedd, L.A. and Hartley, R.J. (1999). *Information seeking in the online age: principles and practice*. London: Bowker-Saur.

Rasmussen, E. (2003). Indexing and retrieval for the web. *Annual Review of Information Science and Technology*, 37, 91-124. Medford, NJ: Information Today Inc.

Salton, G. and McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sullivan, D. (2003). *How search engines rank web pages*. <<http://www.searchenginewatch.com/webmasters/article.php/2167961>>.