

BLOCK 5

INFORMATION RETRIEVAL

THE PEOPLE'S
UNIVERSITY

UNIT 17 INFORMATION RETRIEVAL PROCESSES AND TECHNIQUES

Structure

- 17.0 Objectives
- 17.1 Introduction
- 17.2 Information Retrieval Systems
- 17.3 Databases
 - 17.3.1 Data
 - 17.3.2 The Database
 - 17.3.3 Records and Fields
 - 17.3.4 Properties of Databases
 - 17.3.5 Kinds of Databases
 - 17.3.6 Information Retrieval vs. Database Management Systems
- 17.4 Information Retrieval Systems: Purpose, Components and Functions
- 17.5 Indexing and Information Representation
 - 17.5.1 Inverted File
 - 17.5.2 Access to Inverted Files
- 17.6 Vocabulary Control
 - 17.6.1 Vocabulary Control Tools
 - 17.6.2 Controlled vs. Natural Language Indexing
- 17.7 Searching
 - 17.7.1 Exact Match Search
 - 17.7.2 Best Match Search
 - 17.7.3 Partial Match Search
- 17.8 Information Seeking and User Interfaces
 - 17.8.1 Information Need and Information Seeking
- 17.9 Features of Information Retrieval Systems
 - 17.9.1 Features of Different Types of Information Retrieval Systems
 - 17.9.2 Information Retrieval Features of Online Search Services
 - 17.9.3 Information Retrieval Features of OPACs
 - 17.9.4 Information Retrieval Features of e-Journal Services
 - 17.9.5 Information Retrieval Features of Digital Libraries
- 17.10 Web Information Retrieval Systems
 - 17.10.1 Characteristics of Web Information Retrieval
 - 17.10.2 Information Retrieval Features of Web Search Engines
- 17.11 Intelligent Information Retrieval
 - 17.11.1 Expert Systems for Information Processing and Management Applications
- 17.12 Summary
- 17.13 Answers to Self Check Exercises
- 17.14 Keywords
- 17.15 References and Further Reading

17.0 OBJECTIVES

The aim of any Information Storage and Retrieval (ISAR) system, irrespective of its types, is to retrieve the desired information. Thus, the processes and

techniques of retrieval are of utmost importance. In this Unit, we will be introducing you with the essential features, processes and techniques of information retrieval in non-web and web environments.

After reading this Unit, you will be able to:

- understand the concept of information retrieval (IR) system;
- understand the nature of organisation of records and files which are necessary for retrieval of information particularly the bibliographic information;
- know different types of information retrieval (IR) systems;
- know the information retrieval (IR) processes and techniques in general; and
- understand the features of information retrieval from Web.

17.1 INTRODUCTION

The term information retrieval was coined by Kelvin Mooers over 50 years ago, but it gained popularity in the research community after about a decade, in the early sixties, when computers were being introduced in information handling. The term information retrieval was then being used to mean retrieval of bibliographic information from stored document databases. Truly speaking, these information retrieval systems were document retrieval systems; they were designed to retrieve information about the existence (or non-existence) of bibliographic documents relevant to a user's query. In other words, early information retrieval systems were designed to retrieve an entire document (a book, an article, etc.) in response to a search request. While this is very much what today's information retrieval systems do, over the years many advanced techniques have been developed and applied to design information retrieval systems. These techniques and other modules on information retrieval will be discussed in this Unit. Over the years the connotation of information retrieval has changed and it has been variously termed by information professionals and researchers, some of which include: information storage and retrieval, information organization and retrieval, information processing and retrieval, text retrieval, information representation and retrieval, and information access.

17.2 INFORMATION RETRIEVAL SYSTEMS

An information retrieval system is designed to analyse, process and store sources of information and retrieve those that match a particular user's requirements [Chowdhury, 2004]. Modern information retrieval systems can either retrieve bibliographic items, or the exact text that matches a user's search criteria from a stored database of full texts of documents. Although information retrieval systems originally meant text retrieval systems since they were dealing with textual documents, modern information retrieval systems deal not only with textual information but also with multimedia information comprising text, audio, images and video. While many features of conventional text retrieval systems are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information have called for the development of many new tools and techniques for information retrieval. Thus, modern information retrieval systems deal with storage, organization and access to text, as well as multimedia information resources.

The concept of information retrieval pre-supposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The documents or records that we are concerned with contain bibliographic information which are quite different from other kinds of information or data. We may take a simple example. If we have a database of information

pertaining to an office, or a company, all we have are the different kinds of records and related facts, like names of employees, their positions, salary, and so on, or in the case of a manufacturing company, names of different items, prices, quantity, and so on. The retrieval system here is designed to search for and retrieve specific facts or data, like the salary of a particular manager, or the price of a perfume, and so on. The major objective of an information retrieval system, on the other hand, is to retrieve the information – either the actual information or through the documents containing the information surrogates – that fully or partially match the user's query. Thus, the search output may contain bibliographic details of the documents that matches the query, or the actual text, image, video, etc. that contain the required information. The database in case of an information retrieval system may contain abstracts or full texts of documents, like newspaper articles, handbooks, dictionaries, encyclopedias, legal documents, statistics, etc., as well as audio, images, and video information.

Whatever may be the nature of the database – bibliographic, full-text or multimedia – the system pre-supposes that there is a group of users for whom the system is designed. Users are considered to have certain queries or information needs, and when they put forward their requirement to the system, the latter should be able to provide the necessary bibliographic references of those documents containing either the required information, or the actual text in the case of a full-text retrieval system. Alternative models of (knowledge-based) information retrieval seek to provide the user with the information directly rather than just the citations, the abstract or the full text.

Self Check Exercises

- 1) Mention some synonyms of information retrieval?
- 2) How does the original connotation of IR differ from that of modern 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.

.....

.....

.....

.....

.....

17.3 DATABASES

An information retrieval system deals with databases, and so does a database management system. So, what is the difference between an information retrieval system and a database management system? Before we discuss these differences, we need to have some basic idea of a database and its various components, types, etc. which are discussed in the following sections.

17.3.1 Data

The data is discrete fact, when processed it becomes information. However, in the context of Information Retrieval System, we may consider information as a logical set of data. The word 'data' refers to a set of given facts. Information in a form that can be processed by a computer is called data. The term data has for long been used to refer to scientific measurements, but words also constitute data just as numbers do. A list of names is data, a set of keywords is data, a

doctor's record of his patients is data, and figures relating to temperature, humidity, etc., or sales of a company, are data.

17.3.2 The Database

A database can be conceived as a system whose base, whose key concept, is simply a particular way of handling data. In other words, a database is nothing more than a computer-based record-keeping system. The overall objective of a database is to record and maintain information. The *Macmillan Dictionary of Information Technology* [Longley and Shain, 1989] defines a database as 'a collection of interrelated data stored so that it may be accessed by authorised users with simple user-friendly dialogues'. The *Chambers Science and Technology Dictionary* [Walker, 1988] provides a more simple definition of a database: 'a collection of structured data independent of any particular application'.

It may be noted from the above definitions that a database contains some data that are structured and integrated. Ellingen [1991] defines a database as 'a collection of information that can be searched as a single entity'. According to Oxborrow [1989], a database can be considered as 'an organised collection of related sets of data, managed in such a way as to enable the user or application program to view the complete collection, or a logical subset of the collection, as a single unit'.

From the above definitions we can simplify the definition of a database as an organised collection of related sets of data that can be accessed by more than one user by simple means and can be searched to reveal those that touch upon a particular need [Chowdhury, 2004]. In the computer world we frequently deal with files, which are the outer identifying boundary or a sort of folder containing data. Thus, a file is equivalent to an ordinary address book, if we are talking about names and addresses. A file in a computer is given a unique name by which it is addressed.

17.3.3 Records and Fields

A *record* is a collection of related information. A database is an organised collection of units of information, and each unit of information in a database is called a *record*. A record is generally what a user wants to find out while searching a database. An example of a record is the main entry in a library's catalogue, which describes the book's title, author, subject, etc. A collection of database records constitutes a database file. Identifying what the record is to be is one of the early tasks in designing a database. If the database is a bibliographic one, the bibliographic information about each document is the unit of information or record.

A stored record is a named collection of associated stored fields [Date, 1981]. Each record is made up of particular segments or elements of information, each of which is called a *field*. A field holds a particular type of information within a record that can be separately addressed. The different items of information in a bibliographic record may be author, title, subject heading, etc. Thus, the different fields in a bibliographic record can be the 'author field' containing name(s) of author(s), the 'title field' containing the title, and so on. A field may be subdivided into still smaller units called *subfields*. For example, if 'imprint' of application is regarded as a field in a bibliographic database, then the different components of the imprint – the publisher's name, place of publication, and date of publication – can be called subfields.

A record is, thus, composed of fields and subfields. Identifying what fields and subfields are to be included in each record is an important task in the database

design process. Each field is given a unique identifier, at the design stage, called *field tag*, which is then used throughout for data input, editing, searching, printing, and so on. Several standards have been developed to help the designers of information retrieval systems in this regard. For example, in case of an online catalogue, or more specifically OPAC (Online Public Access Catalogue), as they are called, standard bibliographic format like MARC21 [2002] (MARC stands for Machine Readable Catalogue or Cataloguing; several different types of MARC formats have been developed and MARC21 is the most recent and the most heavily used MARC format), CCF (Common Communication Format) [1992], and so on, specify the fields and the corresponding field tags to be used while preparing catalogue entries for bibliographic items.

17.3.4 Properties of Databases

A database is designed to avoid duplication of data as well as to permit retrieval of information to satisfy a wide variety of user information needs. Major properties of a database can be summarised as follows:

- it is integrated with provisions for different applications;
- it eliminates or reduces data duplication;
- it enhances data independence by permitting application programs to be insensitive to changes in the database;
- it permits shared access;
- it permits finer granularity; and
- it provides facilities for centralised control of accessing and security control functions.

17.3.5 Kinds of Databases

In discussing databases, it is sometimes useful to classify them by the type of data record contained and sometimes by subject coverage. The two major divisions are *reference databases* and *source databases*. Reference databases lead the users to the source of the information: a document, a person or an organisation. They can be divided into three categories:

- a) *bibliographic databases*, which include citations or bibliographic references, and sometimes abstracts of literature;
- b) *catalogue databases*, which show the catalogue of a given library or a group of libraries in a network; and
- c) *referral databases*, which offer references to information such as the name, address, specialisation, etc., of persons, institutions, information systems, etc.

Source databases provide the answer with no need for the user to refer elsewhere. These databases contain the information sought for in electronic form and, therefore, the user can get access to the information instantly as a result of a search. Source databases can be grouped according to their content, for example,

- a) *numeric databases*, which contain numerical data of various kinds, including statistics and survey data.
- b) *full-text databases*, which contain the full text of documents.
- c) *text-numeric databases*, which contain a combination of textual and numerical data, such as a company annual report and handbook data.

Bibliographic databases form the basis of most of the information retrieval systems available today, be they home-grown or available on CD-ROM or through online access. Bibliographic databases can be divided into five broad categories:

- a) large discipline-oriented databases;
- b) interdisciplinary databases with coverage based on key or core journals;
- c) cross-disciplinary databases;
- d) smaller, more specialized databases serving a particular technology or application area; and
- e) databases covering specific types of publication.

However, there could be many more kinds of bibliographic databases, such as:

- *Specific subjects/disciplines*: CAsEarch, BIOSIS, ERIC, MEDLINE, ENERGYLINE, LISA, ISA, and so on;
- *Multidisciplinary*: SCI SEARCH, SOCIAL SCISEARCH;
- *Mission-oriented*: NASA;
- *Problem-oriented*: ENVIROLINE, TOXLINE;
- *Referral*: Foundations Directory, Fine Chemicals Directory, Ulrich's International Periodicals Directory;
- *Factual*: PTS Forecasts, CARIS/FAO (Ongoing Research);
- *Textual references*: DRUGLINE; and so on.

Many of these databases are available online, accessible through the web, and CD-ROM versions.

17.3.6 Information Retrieval vs. Database Management Systems

The technology that helps to process and manipulate data of various kinds is broadly termed as database management technology, and the resulting software packages are known as database management systems (DBMSs). A database management system stores and retrieves discrete data elements that are structured, as opposed to a typical information retrieval system that is designed to deal with unstructured data e.g., the full texts of documents.

Typically a search in a database management environment produces one or more records that are stored in the database. One may argue that an information retrieval system also stores discrete data elements, like author, title, keyword, etc., in the form of a structured database. While this is true, an information retrieval system also handles unstructured data, for example a large chunk of text, and this is where a typical database management system differs from an information retrieval system. Many more differences between the two systems can be noticed especially in the search and retrieval aspects. For example, in a typical database management search, we expect to retrieve discrete data, e.g. the price of an item, date of birth of an employee, and so on, whereas in information retrieval search we retrieve an entire document or part of it containing the information required by the user. The major differences between a typical database management system and an information retrieval system are shown in Table 17.1.

Table 17.1: Difference Between IRS and DBMS

Information Retrieval Systems	Database Management Systems
Designed to deal with unstructured data	Deals with structured data
An item may be retrieved if it exactly or partially exactly matches a query	An item will be retrieved only when it matches the query
Queries are usually language-based, e.g. a keyword, an author name etc.	Queries are mostly value-based, e.g. salary or date of birth of a person
Vocabulary is very important and usually some vocabulary control tools are used	No vocabulary control tool is required
A number of advanced search techniques are used, for example proximity search	Exact match of search term and field value is expected

Self Check Exercises

- 3) What is a database? Give examples of three bibliographic databases.
- 4) Discuss two major differences between a DBMS and an IRS?

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

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

.....

.....

.....

.....

.....

.....

17.4 INFORMATION RETRIEVAL SYSTEMS: PURPOSE, COMPONENTS AND FUNCTIONS

An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user at the right time. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to provide it to the user as soon as asked for. Belkin [1980] presents the following situation which clearly reflects the purpose of information retrieval systems:

- a) a writer presents a set of ideas in a document using a set of concepts;
- b) somewhere there will be some users who require the ideas but may not be able to identify those. In other words, there will be some persons who lack the ideas put forth by the author in his/her work; and
- c) information retrieval systems serve to match the writer's ideas expressed in the document with the users' requirements or demands for those.

Thus, an information retrieval system serves as a bridge between the world of creators or generators of information and the users of that information. The information resources are processed, indexed and stored in an appropriate way. The users interact with the system through a user interface. The user queries, submitted through the interface are matched with the index and the matching items are retrieved. A number of activities are involved in the processes of information processing, indexing and matching. These will be discussed later in this Unit.

As shown in Figure 17.1, the major functions of an information retrieval system can be divided into two categories: (a) organisation and representation of information, and (b) retrieval of information. In the organisation part, although the specific techniques vary from one information retrieval system to another, the basic task is to create an index, called the inverted index, of the potential search terms (keywords and phrases). The index terms may be assigned by a human indexer or by an automatic process, or may be derived automatically from the document texts based on some selection criteria.

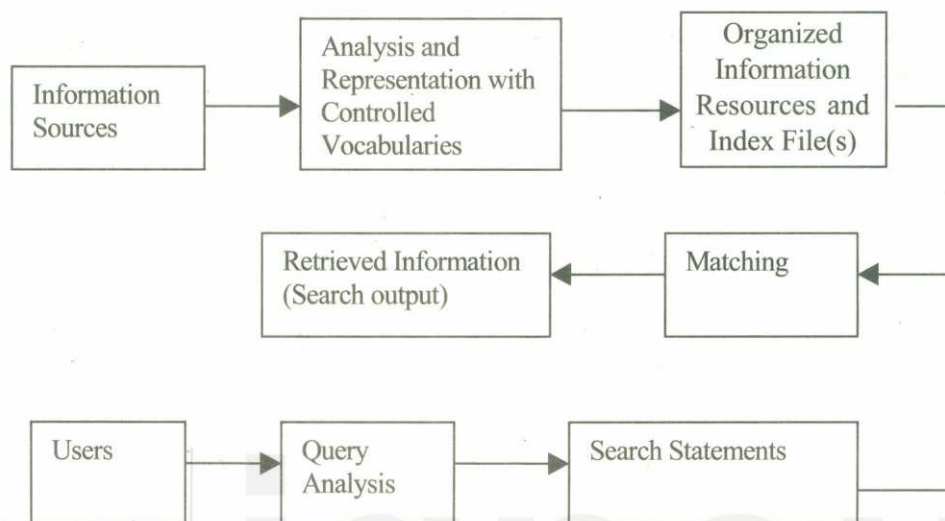


Fig. 17.1: Broad outline of an Information Retrieval System (IRS)

The retrieval process begins with a user query. A user with an information need, interacts with the information retrieval system through the user interface and submits a query. A search query may contain a simple keyword or a phrase, or may contain more than one keyword or phrase combined with some search operators. The retrieval system matches each search term with the inverted index file, and retrieves the matching items. Although this is the basic process in an information retrieval operation, the specifics of each of these activities may be quite complex and depend on the retrieval system, or the retrieval engine, as they are now called. A number of tools may also be used in one or more stages. For example, vocabulary control tools like thesauri, and/or machine translation tools, may be used in the indexing and/or retrieval process.

17.5 INDEXING AND INFORMATION REPRESENTATION

In a typical information retrieval environment, the users queries are not matched with the documents *per se*; instead, they are matched with an index file. The actual documents are stored in a separate sequence, and once a match is found between an index term and a user search term, the pointer from the index file is followed to retrieve the document.

The elementary units of a text retrieval system are the document records. Each document record comprises of a number of fields and subfields, each one of which contains a particular unit of information – author's name, publisher's name, title, keyword(s), class number, ISBN, and so on. The document record may also contain an abstract or full text of the document concerned. A text retrieval system is designed to provide fast access to the records through any of the sought keys or access points. This means that there should be a mechanism for fast access to the document records. What should the basic mechanism be for accessing the document records through some key values – by chosen keyword(s), or by author's name, say? To answer this question, we should first understand how document records are physically stored in the computer.

Document records are stored one after another in the computer memory: this is actually the *virtual structure* of the database file. Imagine a text database that stores a few, say ten, document records. Now, suppose a user wants to check if there is a document in the database that is written by G.G. Chowdhury; another user wants to know if there is a book on Internet. What would the user's approach be? The simplest way would be to open each document record one after another and to check each and every field; if there is a match then it retrieves that document. This process continues until all the document records have been checked. It may be a very simple approach, but one can very well imagine that this will be an extremely slow process even for a faster computer when the text database is relatively large, and will be an impossible proposition for a database that has some hundreds or thousands of document records or more.

What is the solution then? How can we retrieve the desired document records? Let's take a common example. What do we do when we want to locate a particular term or phrase, say the word 'computer', or the phrase 'information retrieval', in a book? Do we start from the first line in the first page and continue up to the last line in the last page of the book? No: we use a simple tool – the back-of-the-book index. What is such an index? It is a simple alphabetical list of all the potential index terms, drawn from the text of the book, each having a pointer showing the occurrence(s) of the terms. Thus, we look into the index file with the required search term, locate it and then move to the page(s) indicated for the actual information. A similar approach is taken in a text retrieval system: an index file is created that contains all the potential index terms arranged in an appropriate order. This index file is called an *inverted file*. Users looking for particular information are required to consult the inverted file first, which then leads to the main database where the document records are stored.

Like inverted file, two more file structures also exist for representation and access to information. These are sequential file and indexed sequential file.

Sequential File: In a sequential file the records are arranged in order of a key field and the computer can use a searching technique, like a binary search, to access a specific record. A sequential file is designed for efficient processing of records in sorted order on some search key. In this file structure, records are chained together by pointers to permit fast retrieval in search key order. Pointer points to next record in order. Records are stored physically in search key order (or as close to this as possible). This minimises number of block accesses.

Indexed Sequential File: Indexed sequential file is a type of file access in which an index is used to obtain the address of the block containing the required record. In indexed sequential files each record of a file has a key field which uniquely identifies that record. It has an index consists of keys and addresses. Indexed sequential files are important for applications where data needs to be accessed either sequentially or randomly using the index.

Example: A library may store details about its users as an indexed sequential file. Sometimes the file is accessed sequentially: when the whole of the file is processed to produce overdue statistics at the end of the month. or

Randomly: may be a user changes address, or a lady user gets married and changes her surname. An indexed sequential file can only be stored on a random access device, e.g., magnetic disc, compact disc (CD).

17.5.1 Inverted File

In an inverted file system of text retrieval, each database consists of two files. One is the text file, which contains what we would expect to find, that is the document records in their normal form – the form in which they are entered into the database. The other is the inverted file, which contains all the index terms, drawn automatically from the document records according to the indexing technique

adopted for the purpose. Each index term in the inverted file is associated with a pointer which shows the record number in which the index term occurs.

The indexing technique, i.e., the technique adopted to draw index terms from the records, determines the order in which index terms will appear in the index file. Different techniques may be required for the purpose: for example, index entries may be required for:

- each and every term occurring in a given field, for example, all the words occurring in the title field. However, there is a risk; some unwanted terms, like 'a', 'an', 'and', 'the', etc., occurring in the document titles may also be indexed. To avoid this problem, text retrieval systems usually incorporate a stop-word file which prevents unwanted terms from being indexed
- the whole field as it is, for example, the full title as it occurs in the document record
- each occurrence of a repeatable field, for example, names of all the authors
- some selected words or phrases from a field or subfield, for example, some terms and phrases occurring in the title field, etc.

Thus, for each significant index term in the database the inverted file contains an entry along with a reference list which specifies position(s) in the database where that term appears. Therefore, in an inverted file system, the searcher first consults the index file, which then refers to the position in the main text database where the desired record appears. The inverted file system is, thus, an example of *indirect file access*. If the terms are arranged alphabetically in the inverted file, then the file represents an example of *indirect sequential file organization*. Figure 17.3 presents a very simple example of such an inverted file which will help us understand the basic concept of an inverted file. However, an inverted file may contain a lot of other information along with each entry, such as the number of occurrences of the term in a given record or position information, such as the field in which the term/phrase occurs, where the term/phrase occurs in a given sentence/paragraph, and so on. As shown in Figure 17.3, index entries are drawn from all four sample document records for the author, title, publisher, and keyword fields. Titles have been indexed as they are, while each occurrence of the author and the keyword field in the document records has been indexed.

Document records

Document no: 1
 Author: Cunningham, M.
 Title: File structure and design
 Publisher: Chartwell-Bratt
 Year: 1985
 Keywords: File structure; File organization

Document no:2
 Author: Tharp, A.
 Title: File organization and processing
 Publisher: John Wiley
 Year: 1988
 Keywords: File structure; File organization

Document no: 3
 Author: Ford, N.
 Title: Expert systems and artificial intelligence
 Publisher: Library Association
 Year: 1991
 Keywords: Expert systems; Artificial intelligence; Knowledge-based systems

Document no: 4
 Author: Charniak, E.; McDermott, D.
 Title: Introduction to artificial intelligence
 Publisher: Addison-Wesley
 Year: 1985

Fig. 17.2: Sample document records

Index file	
4 40 1 1	Addison-Wesley
3 60 1 2	Artificial Intelligence
4 60 1 1	Artificial Intelligence
4 20 1 1	Charniak, E.
1 40 1 1	Chartwell-Bratt
1 20 1 1	Cunningham, M.
3 60 1 1	Expert Systems
4 60 1 2	Expert Systems
3 30 1 1	Expert Systems and Artificial Intelligence
1 60 1 2	File Organization
2 60 1 2	File Organization
2 30 1 1	File Organization and Processing
1 60 1 1	File Structure
2 60 1 1	File Structure
1 30 1 1	File Structure and Design
4 30 1 1	Introduction to Artificial Intelligence
3 60 1 3	Knowledge-based Systems
3 40 1 1	Library Association
4 20 1 2	Mcdermott, D.
2 20 1 1	Tharp, A.

Fig. 17.3: Sample inverted index file

The field tag is used to denote the field where the given term/phrase occurs. This information is used in field-specific searches (discussed in Unit 19). Similarly, the position information is used for proximity or adjacency searching (discussed in Unit 19). Other types of information may also be stored along with each entry, and each such item of information facilitates a particular type of search. Nevertheless, the more such information is added to each entry, the more bulky the inverted file becomes, and therefore takes more storage space and processing time. In this example, a user looking for a term 'expert systems' will retrieve two records, document numbers 3 and 4 from the database, while another user looking for a book written by 'Tharp, A.' will retrieve book number 2. A complex query with search terms combined by Boolean operators will follow the same path. For example, a user with a query 'expert systems OR file organization' will retrieve all four document records, while the query 'artificial intelligence AND knowledge-based systems' will retrieve document record number 3. In the first example, as the search terms are joined by the logical operator 'OR', the system will consult the inverted file for each term and then will merge the document numbers retrieved in each case, while in the second case, because the terms are joined by the logical operator 'AND', the retrieved document numbers for both terms will be matched to locate the common document numbers, that is the ones where both terms are present. Figure 17.3 shows that an index term may occur in several document records, and in each case, several items of information, such as its frequency of occurrence, field(s) in which it has occurred, position information, and so on, have to be stored in the index file. Thus, conceptually the structure of an inverted file may look like the one shown in Figure 17.3.

17.5.2 Access to Inverted Files

The user may pose a single key query or a multiple key query. In the former case, the value of a single search key (say the name of the author) is used as the retrieval criterion, whereas in a multiple key search a number of search keys (say the name of the author, subject name, date of publication, and so on, as in the query 'papers written by Salton on information retrieval systems between

1980 and 1990'). For single key searches, the whole file can be maintained in an order according to the value of the given single set of keys. In a telephone directory, for example, users search through the names of subscribers and therefore the names of subscribers are arranged in alphabetical order. File access in multi-key searches is complicated by the fact that it is not possible to order the file simultaneously in accordance with the values of the different search keys. For example, a users' file in a library can be arranged according to the name of the user, occupation or specialisation, address or department, and so on, and in each case the resulting arrangement of the records within one field will be different from the other.

In the case of a multi-key search, a principal key is to be identified and the file can be ordered in accordance with the values of that key. When the principal key is used as part of a search statement, the subsection of the file corresponding to the given principal key value can then be isolated and subjected to a separate search based on the values of any secondary keys also included in the search query.

A catalogue of a library can be considered as a multi-key file, where the keys are the author, title, publisher, subject, etc. In such a file, the principal key is usually the author, i.e., the file is ordered in accordance with the name (surname) of the authors. From each record in the main file there may be a number of pointers giving access to secondary keys, like publisher, title, etc. A simple file of authors and publishers can be ordered according to the author's name as the principal key, with a sparse index giving access to a chain of pointers for each publisher name. Documents published by a given publisher can be found by following the pointer chain. Pointer chains can be provided for all secondary keys in addition to the primary keys attached to the records; each given record can be traced through the pointer chain for any of the keys. This type of record organisation is known as a multi-list [Chowdhury, 2004].

Multi-list organisation, however, becomes too time-consuming when each query key is attached to a large number of records. One solution to this might be to use large indexes that provide one pointer for each record exhibiting a given key value. Such an index is called an inverted index or an inverted file. Inverted files are widely used in operational information retrieval situations. The advantage of using inverted files is that such files allow extremely rapid search and retrieval operations, based only on the information provided in the index rather than on data from the main record file.

One important issue for the inverted file system is the size of the index file. If each and every term occurring in the document database is indexed, then size of the index file will be quite large, equal to that of the main document database. Therefore, in order to facilitate fast searching, we need to have a method that allows fast access to the terms/phrases in the inverted file. In other words, we need to have an efficient file organization technique.

Self Check Exercises

- 5) What is an inverted file? What role does it play in an information retrieval process?
- 6) What is the difference between a single key and a multi-key query?

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

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

.....

.....

.....

.....

.....

Vocabulary control is one of the most important components of an information retrieval system. As we have noted from its simple model given in Figure 17.1, an information retrieval system tries to match user queries with the stored documents (the inverted index file to be precise) and retrieves those that match. In order to match the contents of the user requirements (the search terms) with the contents of the stored documents (the index entries), one must follow a vocabulary that is common to both. In other words, user requirements need to be translated and put to the retrieval systems in the same language (using the same terms, for example) as was used to express the contents of the document records. This leads us to the concept of using a standard or controlled vocabulary in an information retrieval environment.

Indexing may be thought of as a process of labelling items for future reference. Considerable order can be introduced into the process by standardizing the terms that are to be used as labels. This standardization is known as vocabulary control, the systematic selection of preferred terms.

Lancaster [1986] suggests that the process of subject indexing involves two quite distinct intellectual steps: the 'conceptual analysis' of the documents and 'translation' of the conceptual analysis into a particular vocabulary. The second step in any information retrieval environment involves a 'controlled vocabulary', that is a limited set of terms that must be used to represent the subject matter of documents. Similarly, the process of preparing the search strategy also involves two stages: conceptual analysis and translation. The first step involves an analysis of the request (submitted by the user) to determine what the user is really looking for, and the second step involves translation of the conceptual analysis to the vocabulary of the system.

There are two major objectives of vocabulary control in an information retrieval environment:

- a) to promote the consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials. This is achieved through the control (merging) of synonymous and nearly synonymous expressions and by distinguishing among homographs; and
- b) to facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related.

Lancaster [1986] adds that indexing tends to be more consistent when the vocabulary used is controlled, because indexers are more likely to agree on the terms needed to describe a particular topic if they are selected from a pre-established list than when given a free hand to use any terms they wish. Similarly, from the searcher's point of view, it is easier to identify the terms appropriate to information needs if these terms must be selected from a definitive list. Thus controlled vocabulary tends to match language of indexers and searchers. The various aspects of vocabulary control has been discussed in Unit 2 of this course.

A number of vocabulary control tools have been designed over the years. They differ in their structure and design features, but they all have the same purpose in an information retrieval environment. A number of software packages are now available that allow the record creator to automatically switch to one or more chosen online vocabulary control tools in order to select appropriate terms for representing the document in hand. For example, OCLC's *Connexion* (an integrated cataloguing suite of tools) and OCLC's *CatExpress* (simple copy cataloguing suite of tools) provide such facility. This helps in a number of ways – the document records do not only contain a number of terms that are

representative of the contents of the document, but these are also standardized (in terms of their usage, spelling, form, and so on) and are likely to be chosen by the user for searching purposes. Similarly, there are programs available by which end-users may go to the appropriate page of a particular online vocabulary control tool in order to choose the most appropriate term(s) for preparing the search expression. Vocabulary control tools also help end-users modify their previously formulated search expressions by either widening or narrowing down the search expressions.

17.6.1 Vocabulary Control Tools

As the name suggests, these are the tools used to control the vocabulary of indexing and retrieval. What an indexer and an index user need is a set of guidelines for the proper selection of terms. Syntactic structures are devices that provide these guidelines by showing the relationships among terms or concepts, and they fall into two major categories: (i) classification schemes, and (ii) subject heading lists and thesauri. A combination of the two categories has also been developed.

Classification schemes, being tools for organising knowledge, could be of great help for vocabulary control but the main body of classification schemes is organised in an artificial language (called notations which may contain numbers, alphabets, punctuation marks, or a combination of them) whereas for vocabulary control we need natural language representation. Indexes to classification schemes could serve the role of vocabulary control but here terms appear alphabetically and thus the logical (semantic) organisation of knowledge is not available. Some attempts have been made to combine the features of the main arrangement in classification schemes with those that appear in the index to the classification scheme to generate some kind of faceted or classified thesaurus such as thesauro-facet. Further discussion on these tools are available in Units 2 & 3.

Subject heading lists were initially developed to prepare entries/headings in a subject catalogue that could replicate the classified arrangement of document records. Therefore, they include rather broader subject terms or headings. On the other hand, thesauri have been developed on specific subject fields with a view to bringing together the various representations of terms (synonyms, spelling variants, homonyms, etc.) along with an indication of a mapping of that term in the universe of knowledge by indicating the broader (superordinate), narrower (subordinate), and related (coordinate and collateral) terms. However, this distinction has gradually faded and the latest Library of Congress subject headings list indicates the terms' features as shown in normal thesauri.

17.6.2 Controlled vs. Natural Language Indexing

Controlled indexing languages are those in which both the terms that are used to represent subjects and the process whereby terms are assigned to particular documents are controlled or executed by a person. Normally there is a list of terms - a subject headings list or a thesaurus, that acts as the authority list in identifying terms that may be assigned to documents, and indexing involves the assignation of terms from this list to specific documents. The searcher is expected to consult the same controlled list during formulation of a search strategy. In natural language indexing, any term that appears in the title, abstract or text of a document record may be an index term. There is no mechanism to control the use of terms for such indexing. Similarly, the searcher is not expected to use any controlled list of terms.

Whether to use a controlled vocabulary or to use natural language indexing has been an age-old debate in information retrieval. The major debates in natural versus controlled vocabulary indexing are shown in Table 17.2 [Rowley, 1994; Svenonius, 1986].

<i>Era One</i> –	controlled vocabulary
<i>Era Two</i> –	comparisons of natural and controlled language: major experimental studies noted that natural language can perform as well as controlled vocabulary, but other factors, such as the number of access points, are also significant.
<i>Era Three</i> –	many case studies of limited generalizability. Searching online databases was considered. It was noted that the best performance can be achieved by a combination of controlled and natural language; the number of access points was reaffirmed to have a significant effect; full-text and bibliographic databases were noted to have produced different results.
<i>Era Four</i> –	new advances in user-based systems including OPACs. The value of controlled vocabulary in the context of user-friendly interfaces and the development of knowledge bases were noted.

Aitchison and Gilchrist [2000] provide a detailed comparison of natural and controlled language indexing. The details have been provided in Unit 2 of this course. However, despite much debate extending over more than a century, together with a range of research projects, information scientists have failed to resolve the issue concerning the relative merits and demerits of controlled and natural language. Evidences produced by practice and tested research suggest that controlled language and natural language may be used in conjunction with one another.

Self Check Exercises

- 7) What is the role of a vocabulary control tool in an information retrieval process?
- 8) What is the difference between a subject heading list and a thesaurus from the perspectives of 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.

.....

.....

.....

.....

.....

.....

17.7 SEARCHING

While searching for information in a database, users may approach with some keys. For a bibliographic database, such keys can be author name, title, ISBN, subject keywords, etc. In a non-bibliographic database, these keys will depend on the nature of the database concerned.

In a bibliographic information retrieval environment, searches can be divided into two main classes: known item search and unknown item search. A **known item search** is the one where the user knows something about the item being sought. This may be any key like author, title, publisher, ISBN, and so on. In such a case, user can enter the appropriate key and can get the full details of the item concerned. For example, the user can enter the author name to retrieve the full record. However, very few users actually know about the author, title, etc., of the item that he/she might need at a given instance. Consequently most of the searches are unknown item search. An **unknown item search** is the one where users are not aware of the existence of any document that may solve their

problems. In other words, users do not know whether or not any such item exists that can meet their information requirements. Regarding various aspects of search such as strategy, method and techniques you may refer to Unit 19 of this course.

17.7.1 Exact Match Search

In exact match search, the search engine will only match query terms exactly; it does not allow for truncation, wildcards, or stemming. Exact Match option is nowadays available in Internet-based databases to retrieve more relevant information. Phrase search can be characterized as exact match search, where a phrase is given at the search query that searches whole phrase.

17.7.2 Best Match Search

In best match search, the search engine will match query terms closely, if not exactly. It may allow for truncation, wildcards, or stemming. Best Match search is performed, when exact match could not fetch sufficient number of relevant information. Best match search constructs a tree-structured self-organizing map, where each level of the tree consists of a separate, progressively larger self-organising map. The search for the best match then proceeds level by level, at each time restricting the search to a subset of units that is governed by the location of the best match in the previous, smaller level. The map is taught one level at a time, starting from the smallest level. The best match search can be done even more quickly if the data set is relatively small: the location of the best match in the previous level can be tabulated for each input sample.

17.7.3 Partial Match Search

A partial match is one that matches one or more characters at the end of the text input, but did not match all of the regular expression, although, it may have done so had more input been available. Partial matches are typically used when either validating data input, checking each character as it is entered on the keyboard, or when searching texts that are either too long to load into memory or even into a memory mapped file, or are of indeterminate length, for example the source may be a socket. Some information retrieval systems perform partial match search.

17.8 INFORMATION SEEKING AND USER INTERFACES

The user interface forms an important component of an information retrieval system since it connects the users to the organised information resources. A user interface is the means by which information is transferred between the user and the computer and vice-versa. Well-designed user interfaces should allow the users to better find and fully use the information that the information system provides access to. In fact a good user interface greatly enhances the quality of interactions with information systems [Chowdhury, 2004].

User interfaces basically perform two major functions: (a) they allow users to search or browse an information collection, and (b) they display the results of a search, and often allows users to perform further tasks, like sorting, saving and/or printing the search results, modifying the search query, and so on. The user interface therefore is the most important component of an information retrieval system that a user can see and interact with. The success of an information retrieval system depends significantly on the design and usefulness of the user interface. Hence significant amount of research has taken place in the past few decades on the design, use and evaluation of user interfaces to various kinds of information retrieval systems.

17.8.1 Information Need and Information Seeking

The user is the focal point of all information retrieval systems because the sole objective of any information storage and retrieval system is to transfer information from the source (the database) to the user. Information need is often a vague concept. It is often a result of some unresolved problem(s). It may arise when an individual recognizes that his/her current state of knowledge is insufficient to cope with the task in hand, or to resolve conflicts in a subject area, or to fill a void in some area of knowledge. Information, needed by the user to accomplish a goal – to resolve a problem, to answer a specific question, or to meet a curiosity— may vary from quick and brief information to the most exhaustive and detailed information.

Figure 17.4 shows a simple model of information access. Although it appears to be a very simple model, in essence several complex processes take place throughout the process. Some of these processes are technological and are related to the information retrieval system, users interfaces, etc. Other processes relate to the nature and characteristics of the content as well as the concerned user. The process may take more or less time, and may become simple or complex depending on the nature of the users – their cognitive abilities, background, specific nature of the information need, and so on.

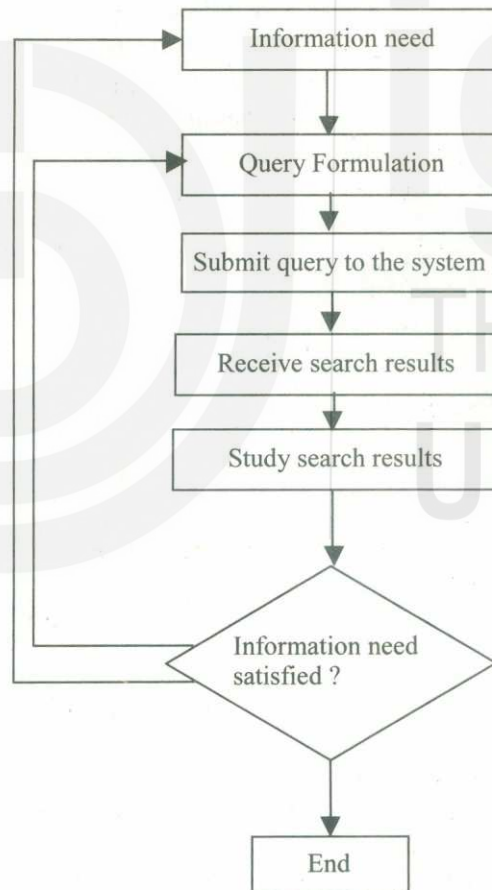


Fig. 17.4: Basic information access model

The basic information seeking and the human information behaviour (HIB) models are based on the model depicted in the above figure.

Self Check Exercises

- 9) What is a user interface?
- 10) What are the two major functions of the user interface in an information retrieval system?

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

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

.....

.....

.....

.....

.....

.....

17.9 FEATURES OF INFORMATION RETRIEVAL SYSTEMS

Based on accessibility of information, two broad categories of information retrieval systems can be identified: in-house and online. In-house information retrieval systems are set up by a particular library or information centre to serve mainly the users within the organization. One particular type of in-house database is the library catalogue. Online public access catalogues (OPACs) provide facilities for library users to carry out online catalogue searches, and then to check the availability of the item required.

By online information retrieval systems we mean those that have been designed to provide access to remote database(s) to a variety of users. Such services are available mostly on a commercial basis, and there are a number of vendors that handle this sort of service. With the development of optical storage technology, another type of information retrieval system appeared on CD-ROM (compact-disc read-only memory). Information retrieval systems based on CD-ROM technology are available mostly on a commercial basis, though there have been some free and in-house developments too. Basic techniques for search and retrieval of information from the in-house or CD-ROM and online information retrieval systems are more or less the same, except that the online system is linked to users at a distance through the electronic communication network.

Recent developments in computer and communication technologies have widened the scope of online information retrieval systems. The Internet and World Wide Web have made information available for use by anyone virtually anywhere with access to the appropriate equipment. This has led to the concept of a digital global library system where information can be generated and made available in electronic form on the Web for use by any user from any corner of the world. This of course involves a number of technical and management issues that need to be resolved in order to make the global digital library concept a reality.

17.9.1 Features of Different Types of Information Retrieval Systems

In today's digital library world, a user can get access to different types of information sources in a variety of formats. For example, a digital library may contain simple catalogues of information resources, like OPACs (Online Public Access Catalogues), or may contain full texts of documents, images, audio and video materials. The information resources may be available in different formats, and they may have been produced by using different types of hardware and software.

For example, the text may be in MS-Word, or PDF, or in HTML format; images may be available in GIF or JPEG file formats. These information resources may reside on a number of different servers – local as well as remote – and they may have been indexed differently. All these issues make the information retrieval process very complex.

The following list represents the common choices that a user may have today from a digital library:

- OPACs
- Electronic databases
 - Online search services
 - CD-ROM databases
- e-Journals
- Digital libraries
 - Local digital libraries
 - Remote digital libraries
- Web resources

Characteristics of the information retrieval systems that work behind all the above information channels or systems are discussed briefly in the following sections.

17.9.2 Information Retrieval Features of Online Search Services

Traditional online information search systems that began about four decades ago were designed to provide access to remote databases, often through a database vendor or service provider. These systems were expensive to use. They were not quite suitable for searching directly by the end-users, and in most cases were used by information intermediaries on behalf of, or in cooperation with, the end-users. Online search services have been provided by database producers, but more commonly by service providers or vendors like *Dialog*, *Ovid*, etc. The major characteristics of this type of online information retrieval system are as follows:

- users get access to remote databases that are often many in number and large in size;
- many databases can be searched using a single search interface;
- database records mainly contain bibliographic details of records with abstracts, and sometimes with additional information, such as citations, etc.; only some databases contain full text information;
- service providers have their own search interface with good search and retrieval capabilities;
- users need to register with the service providers;
- users are charged for searching as well as for the content; and
- modern online service providers have web interfaces with good search features and hyperlinked records/information.

Although each online search service provider, such as *Dialog*, *Ovid*, *STN*, etc., has its own proprietary retrieval engine and user interface, the commonly available search and retrieval features are as follows [Chowdhury and Chowdhury, 2001a]:

- Users can select one or more databases to search;
- Novice and expert search modes are available;

- A search can be conducted with one or more keywords or phrases;
- Common search facilities include: Boolean search, truncation (some systems also allow users to search for the variant forms of a word), proximity search, and field search (number of fields that can be searched depends on the chosen database);
- Searches can be limited by applying certain restrictions, such as language, date, type of material, etc.;
- A search can be conducted for a range of period (date of publication, for example);
- Some systems show the frequency of occurrence of the search terms in the output;
- *Dialog* provides a unique facility of searching through a common index file that allows users to select databases appropriate for a search topic;
- Some systems provide access to thesauri through the search interfaces; and
- Search results can be sorted and sometimes ranked by selected criteria.

17.9.3 Information Retrieval Features of OPACs

Online Public Access Catalogues (OPACs), though are quite different in terms of content, structure etc., from online databases, also provide access to remote databases. OPACs form an important part of a library's services. Features of OPACs can be summarised as follows:

- OPACs allow users to search for the bibliographic records contained within a library's collection;
- Nowadays, some OPACs also provide access to the electronic resources and databases, in addition to the typical bibliographic records;
- Searches take place on the metadata of the records in the library's collection;
- Sometimes users can search more than one collection (within the same library or in different libraries);
- They have relatively simple search interface; and
- OPACs are nowadays available through the web.

Although each OPAC has a search interface and retrieval engine that is proprietary to the company providing the software for the purpose, the following information retrieval features are commonly available in OPACs:

- Browse and search facilities;
- Keyword and phrase search facilities;
- Indexers assign subject headings to the records by using a subject heading list like LCSH (Library of Congress Subject Heading List), and users can search by these assigned subject headings;
- Boolean search usually limited to the keyword search option; in other words, only keywords can be combined with Boolean operators;
- Proximity search also limited to the keyword search option;
- Search results are usually not ranked;
- Searching of records through selected keys – author, title, ISBN, call number, etc.; these are searched as phrases, and are usually automatically right truncated; and
- Some searches can be limited by date, collection, language, etc.

Electronic journals, or e-journals, form a very important part of the collection of today's libraries. Nowadays there are two major categories of e-journals: one that have their printed counterparts, for example, the *Journal of Documentation*, and the other that are available only in electronic format, for example, the *D-Lib Magazine*. Access to electronic journals is provided either by publishers themselves or aggregators. The benefits of getting access to an individual publisher's journals are value-added features and absence of intermediaries. Aggregators, on the other hand, conglomerate journals of several publishers under one interface and search system.

Each publisher and aggregator of e-journals has a proprietary retrieval engine and search interface that can be used to search one or more e-journals. Common information retrieval features of e-journals are:

- users can browse each issue or can search the entire collection;
- there are usually novice and expert search modes;
- word and phrase search facilities are available;
- common search facilities include: Boolean search, truncation, field search, limiting search and range search;
- searches can be conducted on metadata (author, title, etc.) or on the full texts or articles; and
- output is available in one or more formats – HTML, PDF, etc.

17.9.5 Information Retrieval Features of Digital Libraries

Information retrieval services are at the heart of digital libraries [Fox and Urs, 2002]. A digital library can provide access to one or more of the information resources separately using the search interface of each respective system. Alternatively, there may be a single search interface to allow users to conduct searches across all the systems with just one query.

17.9.5.1 Common Features

Based on the study of some selected digital libraries, the following general IR features of digital libraries are observed [Chowdhury and Chowdhury, 2000; Meyyappan, Chowdhury and Foo, 2000]:

- Users can access the collections of a digital library by either of two modes: browsing and searching;
- While most digital libraries allow users to search the local digital library collections, some digital libraries, e.g., NDLTD, provide facilities for federated search or search across a number of digital libraries;
- Boolean, proximity and truncation search facilities are the commonly available search options in the digital libraries, though the operators vary. Some digital libraries, provide options like, 'also must contain', 'or may contain', 'but not contain', 'should contain', 'must contain' and so on, to activate a Boolean search;
- Keyword and phrase search are the common facilities of the digital libraries, though the techniques for conducting a phrase search differs. In some cases, for example in *BUILDER*, users can enter a phrase in the 'Phrase Search Box', while in others, for example, in *DIGILIB* (at University of Queensland, Australia; <http://www.architect.uq.edu.au/digilib/>), *NCSTRL* (Networked Computer Science Technical Reference Library; <http://www.ncstrl.org>), etc., a search phrase has to be entered within double quotes;

- Right truncation and wild card search facilities are common in many digital libraries, and a variety of operators, such as '%', '*', '@', and '?' are used for the purpose. However, some digital libraries provide specific truncation search facilities. For example, in *THOMAS* and *American Memory*, the 'include word variants' option is used for truncation;
- Many digital libraries support proximity search differently. Basically, there are two options: one is through the use of proximity operators, but the operators vary, e.g., 'Near', 'Nearby', 'Sentence', 'Paragraph', and so on;
- Most of the chosen digital libraries allow users to conduct search on specific fields;
- While most digital libraries allow users to specify the maximum number of hits, the output is not always ranked, except for a few like *NDLTD*;
- In some cases, for example, in *ACM* digital library, users can sort the results using some chosen keys; and
- Usually the system comes up with a brief output that can lead to the full records. However, in many cases, an output format can be chosen by the user.

17.9.5.2 Special Features

In addition to the common features mentioned above, some digital libraries have some special information retrieval features. For example,

- *ACM* Digital Library has some unique search features, such as Stem expansion, Fuzzy expansion (spelled like), and sounds like search.
- *DeLiver* (outcome of a DLI1 project at the University of Illinois) offers some unique search features. It allows users to search and view specific parts of the article, such as the figures, or references. Thus user can 'fine tune' a search and get more relevant results.
- *GEMS* (Nanyang Technological University, Singapore, digital library; presently called *iGEMS*) allows users to set up his/her own profile for future search and for obtaining SDI services. It also allows instant opening of a CD-ROM and provides access to an online journal or database.
- *HEADLINE* (a hybrid library project in the UK, <http://www.headline.ac.uk>) is unique in two respects:
 - It automatically creates an information page, called the Subject Page, on the subject of interest of the user. The necessary information is gathered from the user's log-in screen.
 - Allows user to customize the Subject Page to create his/her own subject page.
- *IEL* (IEEE Electronic Library) allows users to choose options to match similar subjects, or to search for the latest additions to IEL. Search interface allows to browse and select search terms from the displayed list. Superscript, subscript and special characters can be searched.
- New Zealand Digital Library has developed a digital library software and makes it freely available, i.e., Greenstone Digital Library Software (GSDL).
- *NDLTD* uses the *InfoSeek* search engine, and therefore a number of good search features are available. Users can search a specific site search or can conduct a federated Search across the digital libraries that are member of the NDLTD Federation.
- *THOMAS* uses a *probabilistic information retrieval system* called 'InQuery'.
- The UC Berkeley DL uses *Cha-Cha* and *CheshireII* search systems and has two unique features:

- Natural language search facilities
- Image retrieval by image content
- *The Universal Library* (at Carnegie-Mellon University; <http://www.ul.cs.cmu.edu/>) has a unique feature called the *hyperbolic tree* that has a unique visualisation effect and user search the collection through this hyperbolic tree.

17.10 WEB INFORMATION RETRIEVAL SYSTEMS

Web information retrieval is significantly different from the traditional text retrieval systems. These differences are mainly due to a number of typical characteristics of the world wide web such as the distributed architecture of the web, the variety of information available on the web, growth of the web, the distribution of information as well as the users, and so on. Major characteristics of the web that make Internet information retrieval different from other information retrieval systems are discussed below.

17.10.1 Characteristics of Web Information Retrieval

- ***Distributed nature of the web:*** The web resources are distributed all over the world. Hence complex measures are required to locate, index and retrieve the information resources. The fact that the computers that are interconnected have different architecture, and the information resources are created using different platforms, software and standards, make the matter more complex. Most text retrieval systems deal with a set of information resources that is several times smaller in volume compared to the web. In addition, text retrieval systems usually deal with a set of documents that have been created using a set of standards – hardware, software and processing standards. Although information retrieval in case of OPACs has to deal with distributed information, the problems are tackled by use of several standards for processing information, such as the MARC formats. No such uniform standard is used for the creation and processing of web information resources.
- ***Size and growth of the web:*** The web has grown exponentially over the past decade. The processes of identifying, indexing and retrieving information become more complex as the size of the web, and hence the volume of information on the web, increases. Conventional text retrieval systems have to be tested and modified to make them suitable for handling the large volume of data on the web.
- ***Deep vs. the surface web:*** Information resources on the web can be accessed at two different levels. While millions of web information resources can be accessed by anyone, a lot of information is accessible either through authorised access (information that is password protected, say) or can be generated only by activating an appropriate program. Researchers call the former as the surface web and the latter the deep web, with a note that the deep web is several times larger than the surface web.
- ***Type and format of the documents:*** While text retrieval systems deal with textual information, the web contains from simple text to multimedia information. Again these information resources appear in a variety of formats thereby making the task of indexing and retrieval more complex.
- ***Quality of information:*** Since anyone can publish almost anything on the web, it is very difficult to assess the quality of the information resources. As opposed to the conventional text retrieval systems that deal with published information resources which are somewhat quality-controlled, web information retrieval systems have to deal with both the controlled and uncontrolled sets of information resources.
- ***Frequency of changes:*** Web pages change quite frequently. This is in sharp contrast with the input of the conventional text retrieval systems that deal with relatively static information; once an information resource is added to a text retrieval system,

it does not change its content, at the most the entire document is removed from the system. Keeping track of the changes in the millions of web pages, and making necessary changes in the information retrieval system is a major challenge. Another major problem with the web is that the resources (the web pages) often move. This information needs to be tracked by the retrieval system in order to facilitate proper retrieval.

- **Ownership:** Information resources that are accessible through the web have different access requirements; while some information can be accessed and used for free, others require specific permission or access rights, often through payment of fees. Identifying the rights to access is a major challenge for web information retrieval.
- **Distributed users:** Most text retrieval systems are designed to meet the information needs of a specific user community. Hence text retrieval systems usually have an idea of the nature, characteristics, information needs, search behaviour etc., of the target user community. Web information is in sharp contrast with this. Ideally the users of an information resource on the web may be anyone, located anywhere in the world. This imposes a significant challenge since the designer of a web information retrieval system will have no idea about the target users, their nature, characteristics, location, information search behaviour, etc.
- **Multiple languages:** Since the web is distributed all over the world, the language of the information resources as well as the users vary significantly. An ideal web information retrieval system should be able to retrieve the required information irrespective of the language of the query or the source information. This diversity of language poses a tremendous challenge for web information retrieval.
- **Resource requirements:** Massive amount of resources are required to build and run an effective and efficient web information retrieval system. The matter is worsened by the fact that there is no single body who would fund for these resources, and yet everyone wants a good information retrieval system for access to web information resources.

17.10.2 Information Retrieval Features of Web Search Engines

Search engines are the most commonly used tools for finding information on the web. Digital libraries usually provide links to one or more search engines to allow users search for the web information resources. A search engine allows the user to enter search terms – keywords and/or phrases – that are run against a database containing information on the web pages collected automatically by programs called Spiders. At the end of a search session, the search engine retrieves web pages from its database that match the search terms entered by the searcher.

There are three main components of a search engine: (1) the Spider, i.e., the program that automatically collects information about millions of pages on the web, (2) the Index that stores information collected by the spider on the various web pages, and (3) the Search engine software and interface with which the users interact to conduct a web search [Chowdhury and Chowdhury, 2001b]. Search engines can be categorised in a number of ways. Two broad categories are: search engines and meta search engines, the later category refers to tools that allow users to conduct concurrent searches on more than one search engines. Some people also categorise search engines based on their characteristics of indexing. For example, Search engines can be categorised as full-text search tools, extracting search tools, subject-specific search tools and meta search tools. *SearchEngineWatch.com*, the most up-to-date and widely used information resource on web search engines, categorises search engines as follows [Sullivan, 2004]:

- The Major Search Engines, e.g., AltaVista, AOL Search, Google, HotBot, etc.
- Kids Search Engines, e.g., Yahoo!igans, KidsClick, etc.
- News Search Engines, AltaVista News, Ananova, Yahoo News, etc.
- Specialty Search Engines, AskJeeves, Allexperts.com, CNETDownload.com, etc.
- Multimedia Search Engines, e.g., AltaVista Photofinder, FAST multimedia search, Ditto, Napster, Gnutella, etc.
- Search Utilities, e.g., Copernic, LexiBot, SearchWolf, Subject Search Spider, etc.
- Paid Listings Search Engines, e.g., Google AdWords, FindWhat.com, Espotting.com, etc.
- Metacrawlers, e.g., Kartoo, Query Server, Profusion, InfoGrid, etc.
- Regional Search Engines, e.g., Mosaique, Indiainfo.com, etc.

Each search engine has a proprietary software for all its information storage and retrieval operations. Consequently each search engine has a specific set of search and retrieval features. The following are the common information retrieval facilities provided by most search engines:

- Word, phrase and natural language search facilities are available;
- Simple and Advanced Search options are available;
- Special options are available for image, audio and video search;
- A number of categories are available for browsing;
- Most search engines support multilingual search;
- Boolean Search: AND, OR, AND NOT, and – signs are used to indicate that the following search term must and must not appear; Parentheses can be used for nested Boolean search. In the Advanced search mode there may be options like: Must Have, Good to Have and Must Not Have;
- Advanced search facilities include: proximity search, truncation, meta tag search (field search);
- A search can be constrained by different criteria, for example, by language, date of publication, collection type, etc.;
- The search output is ranked based on some criteria set by the search engine software;
- Some of the advanced search facilities include the following:
 - *Link*: The keyword 'link:' followed by a domain name or a complete URL (Uniform Resource Locator; simply speaking the address of a web page) returns every Web page that has a hypertext link to a particular site, directory, or page (available in AltaVista).
 - *Translate*: Automatic translation of web pages from selected languages is available. Some search engines also allow users to enter text in a given language which can be instantly translated into another chosen language.
 - *Family Filter*: Can be turned on or off to allow/avoid retrieval of unwanted materials.

Self Check Exercises

- 11) What is meant by online information retrieval?
- 12) How does an OPAC search differ from an online search?
- 13) What is a web search engine and what role does it play in the context of the web?

14) How does the traditional online information retrieval differ from web information retrieval?

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

.....

.....

.....

.....

.....

.....

.....

17.11 INTELLIGENT INFORMATION RETRIEVAL

Sparck Jones [1983] defines an intelligent information retrieval system as a computer system with inferential capabilities such that it can use prior knowledge to establish a connection between a user's (probably ill-specified) request and a candidate set of relevant documents. According to Brooks [1987] an intelligent information retrieval system is a system that carries out intelligent retrieval. Brooks further defines intelligent retrieval as the use, by a computer system, of the stored knowledge of its world of documents, users, etc. and of information about the user and his/her problem to infer which documents would enable that particular user to resolve or manage his/her problem in a better way. It has become apparent through different experiments that retrieval cannot be carried out intelligently unless the system 'knows' about its task, world of documents, language, subject domains, etc., as well as the specific requirement of the user. The realization of the need to use knowledge within retrieval systems has led researchers to look at the disciplines of artificial intelligence and expert systems that also aim to incorporate and use knowledge.

17.11.1 Expert Systems for Information Processing and Management Applications

Over the years researchers have developed several expert systems for professional tasks in both traditional and non-traditional library and information services and management. These tasks include: indexing, abstracting, thesaurus construction, cataloguing and classification, Boolean text retrieval, non-Boolean text retrieval including reference services, automatic content analysis and knowledge representation, relational database access and management, intelligent documents, training, database selection, and database analysis.

Lancaster and Warner [2001] provide an excellent review of the applications of expert systems and related intelligent technologies in different areas of library and information science. They note that the major applications of intelligent technologies in the field of library and information science include the following:

- cataloguing
- subject indexing
- collection management
- reference services including:
 - referral of users to appropriate information resources
 - selection of an appropriate database for searching information to meet a specific information need

- text processing including:
 - text categorisation
 - text summarisation
 - intelligent agents for text processing
 - text mining, data mining and knowledge discovery
- user interfaces.

Some of the applications of intelligent techniques mentioned above, such as in the area of collection management, must have in some form or another, intelligent techniques for processing and retrieval of information for specific tasks.

Self Check Exercises

- 15) What is meant by intelligent information retrieval?
- 16) How does intelligent information retrieval differ from a conventional information retrieval?

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

.....

.....

.....

.....

.....

.....

17.12 SUMMARY

This Unit is designed to provide the basic concept of information retrieval including the processes and techniques involved in the process. Early information retrieval systems were designed to deal with bibliographic information and would allow users to search the collection using some keys like author name, keyword, etc.

With the advent of online information retrieval systems, search techniques and user interfaces became more sophisticated. Side by side use of vocabulary control tools for indexing and searching bibliographic databases became more and more common. The same practices were followed in case of CD-ROM databases.

However, over time the connotation of information retrieval also changed and first came full text and then multimedia information retrieval that called for more sophisticated indexing and retrieval techniques. Although the basic nature of information retrieval remained the same, in the sense that retrieval was done through the inverted index files, more sophisticated techniques for the creation and organization of index files were developed.

The nature of information access and retrieval changed drastically over the past ten or so years with the advent of the World Wide Web and digital libraries. Web search engines brought significant changes in terms of sophistication in indexing and searching. As discussed in this module, users in today's web and digital library world can get access to information through a number of channels, and the underlying information retrieval systems vary significantly in terms of the features and facilities.

Moreover, many more challenges are now faced by today's information

professionals as far as the information processing and retrieval activities are concerned. Over the years many information retrieval models have been developed that provide guidelines for automatic indexing and searching information resources. Both the system and user-centred models of information retrieval have been discussed in Unit 18, while the search strategies, and advanced information search techniques have been discussed in Unit 19.

17.13 ANSWERS TO SELF CHECK EXERCISES

- 1) Information retrieval is also called by: information storage and retrieval, information organisation and retrieval, information processing and retrieval, text retrieval, information representation and retrieval, and information access.
- 2) Originally the term information retrieval was used to mean retrieval of bibliographic information from stored document databases. Since they were primarily dealing with text databases, they were also called text retrieval systems. Modern information retrieval systems deal with multimedia information comprising text, audio, images and video. While many features of conventional text retrieval systems are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information have called for the development of many new tools and techniques for information retrieval. Modern information retrieval systems deal with storage, organisation and access to text, as well as multimedia information resources.
- 3) A database is an organised collection of related sets of data that can be accessed by more than one user by simple means and can be searched to reveal those that touch upon a particular need. Examples of bibliographic databases in clued: Medline, ERIC, BIOSIS, LISA, ISA, etc.
- 4) A database management system stores and retrieves discrete data elements that are structured, as opposed to a typical information retrieval system that is designed to deal with unstructured data e.g., the full texts of documents. In a typical database management search, we expect to retrieve discrete data, e.g. the price of an item, date of birth of an employee, and so on, whereas in information retrieval search we retrieve an entire document or part of it containing the information required by the user.
- 5) An inverted file is an index file that contains all the index terms, drawn automatically from the document records according to the indexing technique adopted for the purpose. Each index term in the inverted file is associated with a pointer that specifies position(s) in the database where that term appears. Therefore, in an inverted file system, the searcher first consults the index file, which then refers to the position in the main text database where the desired record appears.
- 6) In a single key search, the value of a single search key (say the name of the author) is used as the retrieval criterion, whereas in a multiple key search a number of search keys (say the name of the author, subject name, date of publication, and so on). An example of a multi-key search could be: 'papers written by Salton on information retrieval systems between 1980 and 1990'. For single key searches, the whole file can be maintained in an order according to the value of the given single set of keys. In a telephone directory, for example, users search through the names of subscribers and therefore the names of subscribers are arranged in alphabetical order. File access in multi-key searches is complicated by the fact that it is not possible to order the file simultaneously in accordance with the values of the different search keys.
- 7) There are two major objectives of vocabulary control in an information retrieval environment:
 - to promote the consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials.

- to facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related.
- 8) Subject heading lists were initially developed to prepare entries/headings in a subject catalogue that could replicate the classified arrangement of document records. Therefore, they include rather broader subject terms or headings, and they are used primarily to assign subject headings for bibliographic items in a library catalogue or index. On the other hand thesauri have been developed on specific subject fields with a view to bringing together the various representations of terms (synonyms, spelling variants, homonyms, etc.) along with an indication of a mapping of that term in the universe of knowledge by indicating the broader (super-ordinate), narrower (subordinate), and related (coordinate and collateral) terms. Thesauri are now used for indexing and searching electronic databases.
 - 9) A user interface is the means by which information is transferred between the user and the computer and vice-versa. It is an important component of an information retrieval system since it connects the users to the organized information resources.
 - 10) User interfaces basically perform two major functions: (1) they allow users to search or browse an information collection, and (2) they display the results of a search, and often allows users to perform further tasks, like sorting, saving and/or printing the search results, modifying the search query, etc.
 - 11) By online information retrieval systems we mean those that have been designed to provide access to remote database(s) to a variety of users. Such services appeared about four decades ago, and are available mostly on a commercial basis, and there are a number of vendors that handle this sort of service. A prominent example of online search service provider is Dialog. However, in today's world many people use the term online information retrieval to mean retrieval of information from remote databases (and such services are not necessarily restricted to those that are provided by vendors like Dialog).
 - 12) An OPAC search is conducted on a given library's catalogue. Therefore the search results are expected to be in the collection of the given library. An online search on the hand does not have any such restrictions, i.e., the search output is not restricted to any library's collection. The search features in an OPAC environment are much more limited compared to an online search environment. An OPAC search is conducted on the index file created by selecting data from some selected fields in a library's catalogue like author, subject heading, keywords, title, ISBN, etc. However, in case of an online search the search terms may come from the bibliographic details as well as the abstract or the full text of the documents.
 - 13) Search engines are the most commonly used tools for finding information on the web. There are three main components of a search engine: (1) the spider, i.e., the program that automatically collects information about millions of pages on the web, (2) the index that stores information collected by the spider on the various web pages, and (3) the search engine software and interface with which the users interact to conduct a web search.
 - 14) Traditional online information retrieval differs from web information retrieval in a number of way. For example, online information retrieval is proprietary; one has to register and pay for the service, while web information retrieval through web search engines is available to everyone for free. However, online search services guarantee quality information because they follow specific selection criteria for selection of input resources, whereas web search engines do not have such a quality control, and therefore the search output may not always be of high quality. Web search engines are continuously evolving and they offer many sophisticated retrieval facilities that are not available in online search services.
 - 15) There are several definitions of intelligent information and one can choose anyone:

- 1) A computer system with inferential capabilities such that it can use prior knowledge to establish a connection between a user's request and a candidate set of relevant documents.
 - 2) A system that carries out intelligent retrieval. An intelligent retrieval is defined as the use, by a computer system, of the stored knowledge of its world of documents, users, etc. and of information about the user and his/her problem to infer which documents would enable that particular user to resolve or manage his/her problem in a better way.
- 16) A conventional information retrieval system is based on matching the search and the index terms. So, it retrieves items that match the search terms. Since user search terms and index terms may differ in a number of ways, information conventional information retrieval systems often fail to produce the desired results – they may produce unwanted materials, or may fail to produce the relevant materials.

Intelligent information retrieval systems use intelligent techniques such as those used in artificial intelligence and expert systems field, as well as natural language processing techniques to overcome the problems of conventional information retrieval systems that suffer from term matching. Such systems aim to use expert techniques, in the line of the actions of human experts, in finding information relevant to a user need. As a result they are more resource intensive, they are very expensive to build and maintain, and often have applications in a limited subject domain.

17.14 KEYWORDS

Data Mining

: The process of analysing data to identify patterns or relationships. For example, a data mining program might analyze millions of document supply orders to determine trends among top-requesting customers, such as their likelihood to obtain again, or their likelihood to switch to a different vendor.

Database

: An organised collection of related sets of data that can be accessed by more than one user by simple means and can be searched to reveal those that touch upon a particular need.

DLI

: Digital Library Initiative.

Expert System

: A computer system that embodies knowledge about a specific problem domain and can solve problems from the domain using its knowledge with a degree of expertise that is comparable to that of a human expert.

Intelligent Information

: A system that carries out intelligent retrieval, i.e., the use,

Retrieval System

by a computer system, of the stored knowledge of its world of documents, users, etc. and of information about the user and his/her problem to infer which documents would enable that particular user to resolve or manage his/her problem in a better way.

NDLTD

: The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organisation dedicated to promotion and dissemination of electronic theses and dissertations.

: A tool that is used to conduct search in an information retrieval environment. The term is typically used in the context of web information retrieval where a web search engine is a program that allows the user to enter search terms – keywords and/or phrases – that are run against a database containing information on the web pages collected automatically by programs called spiders. At the end of a search session, the search engine retrieves web pages from its database that match the search terms entered by the searcher.

User Interface

: The contact point or the means by which information is transferred between the user and the computer and vice-versa.

17.15 REFERENCES AND FURTHER READING

Aitchison, J. and Gilchrist, A. (2000). *Thesaurus construction and use: a practical manual*. 4th ed. London: Aslib.

Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5, 133–43.

Brooks, H.M. (1987). Expert systems and intelligent information retrieval. *Information Processing and Management*, 23 (4), 367–82.

BS 5723:1987. *Guidelines for the establishment and development of monolingual thesauri*. London: British Standards Institution.

BS 6723:1985. *Guidelines for the establishment and development of multilingual thesauri*. London: British Standards Institution.

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

Chowdhury, G.G. and Chowdhury, S. (2000). An overview of the information retrieval features of twenty digital libraries. *Program*, 34(4), 341–73.

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

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

Date, C.J. (1981). *An introduction to database systems*. 3rd ed. Reading, MA: Addison-Wesley.

Ellingen, D.C. (1991). Database design. *Database*, 14(3), 104–6.

Fox, E. A. and Urs, S. (2002). Digital libraries. *Annual Review of Information Science and Technology*, 36, 503–89. Medford, NJ: Information Today Inc.

Gopinath M.A. (2004). *Multiple database searching and common command language*. In: MLIS-03 Unit 16 course material. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *Information retrieval process*. In: MLIS-03, Unit 13 course materials. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *ISAR systems: operations and design*. In: MLIS-03, Unit 10 course materials. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *Objectives of information storage and retrieval systems*. In: MLIS-03, Unit 16 course materials. New Delhi: Indira Gandhi National Open University.

- Lancaster, F.W. (1986). *Vocabulary control for information retrieval*. 2nd ed. Arlington, VA: Information Resources.
- Lancaster, F.W. and Warner, A. (2001). *Intelligent technologies in library and information science applications*. Medford, NJ: Information Today Inc.
- Longley, D. and Shain, M. (eds.). (1989). *Macmillan Dictionary of Information Technology*. 3rd ed. London: Macmillan.
- Meyyappan, N., Chowdhury, G.G. and Foo, S. (2000). A review of twenty digital libraries. *Journal of Information Science*, 26(5), 331–48.
- Oxborrow, E. (1989). *Databases and database management systems: concepts and issues*. 2nd ed. Chichester: Chartwell-Bratt.
- Rowley, J.E. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), 108–19.
- Salton, G. and McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Sparck Jones, K. (1983). *Intelligent retrieval*. In: *Intelligent information retrieval: Informatics 7*, edited by K.P. Jones. pp. 136–42. London: Aslib.
- Sullivan, Danny (ed.). (2005). *The Search Engine Report*. <<http://searchenginewatch.com/sereport>>.
- Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5), 331–40.
- Walker, P.M.B. (1988). *Chambers Science and Technology Dictionary*. Cambridge: Chambers.

THE PEOPLE'S
UNIVERSITY

UNIT 18 INFORMATION RETRIEVAL MODELS AND THEIR APPLICATIONS

Structure

- 18.0 Objectives
- 18.1 Introduction
- 18.2 Information Retrieval
 - 18.2.1 Database
 - 18.2.2 Information Base
 - 18.2.3 Structured Query Language (SQL)
- 18.3 Information Retrieval Techniques
- 18.4 Models Based on Input/Output
 - 18.4.1 Data Retrieval Model
 - 18.4.2 Information Retrieval Model
 - 18.4.3 Knowledge Retrieval Model
- 18.5 Models Based on Theories and Tools
 - 18.5.1 Boolean Retrieval Model
 - 18.5.2 Fuzzy Logic Model
 - 18.5.3 Set Theoretic Model
 - 18.5.4 Vector Space Model
 - 18.5.5 Probabilistic Retrieval Model
 - 18.5.6 Linguistic Model
 - 18.5.7 Mathematical Model
 - 18.5.8 Psychological Model
 - 18.5.9 Economic Model
 - 18.5.10 Hypertext Linkage Model
- 18.6 Summary
- 18.7 Answers to Self Check Exercises
- 18.8 Keywords
- 18.9 References and Further Reading

18.0 OBJECTIVES

In the efforts for search for appropriate techniques of information retrieval, various models have been developed. This Unit deals with basics of information retrieval techniques and the different types of models developed from time to time. After reading this Unit, you will be able to:

- know the basics and types of factors involved in the information retrieval;
- know the shift from the conventional to modern information retrieval;
- understand the process of matching information need and retrieval of information from databases, knowledge bases, information systems and libraries;
- be conversant with the development of information retrieval models; and
- acquaint yourself with research areas in the field of information retrieval.

18.1 INTRODUCTION

Pattern of information retrieval indicates a knowledge seeking behaviour of individuals and groups of individuals. In this context, a searcher seeks some information from the vast store of a knowledge. An analysis and diagnosis of this state of mind provides guidelines for organisation of knowledge in libraries, information retrieval systems, databases, knowledge bases and similar environments. Such guidelines are aimed at providing conducive compatibility between searchers' approach and knowledge organisation in a database. The current human environment consisting of learning, problem-solving, and decision-making situation calls for flexibility in knowledge structures at each instance. The development in computer and communication technologies have made it possible to store vast amount of information in compact form. The variety of software developed have also given scope for quick retrieval of information from this store. While the speed of retrieval is valuable, it could be enriched further if the retrieved information is readily assimilable by the information seeker. It is in this context that modelling of retrieved information into user-friendly approaches calls for cognitive modelling of information retrieval.

Such development has given rise to a field 'Cognitive Science' which is an interdisciplinary field drawing inputs from the fields of Psychology, Behavioural Studies, Computer Science, Artificial Intelligence and Information Science.

18.2 INFORMATION RETRIEVAL

Information Retrieval is a process of selecting information from a store. It connotes that search for information may be from documents, metadata which describes documents or searching interim databases. The databases may be standalone databases or hypertext networked databases like Intranet and Internet. It primarily helps a person who needs to get some information in his activities, be it research, problem solving, decision-making, production, service, etc. Broadly speaking, four kinds of information retrieval exist. *Document retrieval*, in which simple structured files are normally processed, using a small number of well-defined attributes to characterise each record, and a restricted set of pre-specified query types to access the database; *reference retrieval*, in which the records represent books, documents, and other library materials, and the number of different attributes available for the identification of the information items is effectively unlimited. In that case, the queries often refer to the information content of individual documents. In the most general case, a retrieval system might be designed to handle any kind of query and the system might furnish direct replies to such queries; in *fact retrieval*, a wide variety of different types of information identifiers may be needed, and the answers may have to be based not only on a deep analysis of each individual information item, but also on general world knowledge and other extraneous factors. In *text retrieval*, instead of retrieving reference or data or surrogates of documents, text on a particular topic are retrieved. Irrespective of any retrieval environment, the following four main system components must be taken into account in formulation of the retrieval problem.

- a) **The objects, documents, or records themselves** (which in the aggregate constitute the information files to be processed);
- b) **The information identifiers, terms, index terms, keywords, attributes, etc.** (which characterise the records or documents and represent the information content in each case);
- c) **The information requests** (which enter into the system and are to be compared with the stored records for retrieval); and

- d) **The relevance information** (often supplied by the users of the system connecting the information requests to the stored information items).

An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multimedia objects. An Information Retrieval System thus consists of a software program that facilitates a user in finding the information of his needs. The system may use standard computer hardware or specialized hardware to support the search sub-function and to convert non-textual sources to a searchable media (e.g., transcription of audio to text). The success of an information system is gauged by how well it can minimize the overhead for a user to find the needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select items to read, reading non-relevant items). The success of an information system is very subjective, based upon what information is needed and the willingness of a user to accept overhead. Under some circumstances, needed information can be defined as all information that is in the system that relates to a user's need. Thus, search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead.

The first information retrieval system originated with the need to organize information in libraries. Catalogues were created to facilitate the identification and retrieval of items. The term 'item' is used to represent the smallest complete unit that is processed and manipulated by the system. The definition of item varies by how a specific source treats information. A complete document, such as a book, newspaper or magazine could be an item. Each chapter, or article may also be defined as an item. As sources vary and systems include more complex processing, an item may address even lower levels of abstraction such as a contiguous passage of text or a paragraph.

With the advent of inexpensive powerful personal computer processing systems and high speed, large capacity secondary storage products, it has become commercially feasible to provide large textual information databases for the average user.

In information retrieval the term 'relevant' item is used to represent an item containing the needed information. From a user's perspective 'relevant' and 'needed' are synonymous. From a system perspective, information could be relevant to a search statement (i.e., matching the criteria of the search statement) even though it is not needed, relevant to user (e.g., the user already knew the information).

In addition to the complexities in generating a query, quite often the user is not an expert in the area that is being searched and lacks domain specific vocabulary that is unique to that particular subject area. The user starts the search process with a general concept of the information required, but does not have a focused definition of exactly what is needed. A limited knowledge of the vocabulary associated with a particular area along with lack of focus on exactly what information is needed leads to use of inaccurate and in some cases misleading search terms. Even when the user is an expert in the area being searched, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary.

There are natural obstacles to specification of the information, a user needs that come from ambiguities inherent in languages, limits to the user's ability to express what information is needed and differences between the user's vocabulary corpus and that of the authors of the items in the database. Languages suffer from word ambiguities such as homographs and use of acronyms that allow the same word

to have multiple meanings. Many users have trouble in generating a good search statement. The typical user does not have significant experience with, nor even the aptitude for Boolean logic statements. The use of Boolean logic is a legacy from the evolution of database management systems. Multi-media also adds an additional level of complexity in search specification.

Thus, an information retrieval system must provide tools to help to overcome the search specification problems. In particular, the search tools must assist the user automatically and through system interaction in developing a search specification that represents the need of the user and the writing style of diverse authors.

There are three basic components in a generalized modern computer-based information retrieval system. They are, the database; the information seeker; and the retrieval techniques, tools, models, and processes which attempt to bridge the gap between searcher and the database.

18.2.1 Database

It is designed to take-in information and information sources acquired for the specific purpose of serving user groups. The selection of information and other sources is based on this objective of serving them. A database has a logical and physical organisation. It is an arrangement based on the users' approach. The main purpose of a bibliographic database is to collect and collate standard bibliographic information, assign index terms and abstracts for technical papers and monographs, etc. In case of full text databases, the complete text of document is made available along with bibliographic details.

Information Retrieval Systems are increasingly using unformatted, free form means of storing information. Record-keeping files in a traditional paper based system might have, for example, a separate line for a person's name, a separate line for an address, a place for a seven-digit phone number, and so on. In computer systems, it is now more common to have information stored in a less specifically located form, making more difficult to locate specific features such as a name or an address. The increased use of word processing, the proliferation of desktop personal computers, and the introduction of optical scanners that can read and convert material into organizational databases, which store information in a way that can make it more difficult to recreate structurally, are making information storage systems more popular.

On the other hand, precisely because information can be stored in a free form, unformatted way, electronic means of storage also make it possible to retrieve and manipulate information in ways that are not possible in a paper-based system, as our increasing reliance on electronic means of storage and retrieval system makes it clear. The difference in power and flexibility in retrieving information can be illustrated by comparing searches done in print indexes and their electronic counterparts. A search in a print index restricts the user to chosen keywords or index terms that can be combined in relatively limited ways. A search in an electronic database generally allows freedom to the searcher and allows him to define more precisely the hierarchical and proximal relationships the terms should have.

18.2.2 Information Base

An Information Retrieval System locates and presents information to the user, based on a query presented to the system. A query may indicate precisely the characteristics of information to be retrieved, or it may express as approximation of information need, indicating merely an initial guess as to the characteristics of the information to be retrieved. A precise-query is most commonly found where an answer is needed to a single factual question, such as "which is the capital of Gujarat State". Question requiring more complex or more ambiguous answers may be expressed by queries that are less precise. Take for instance, an

organization that desires to hire internally to establish an information system that will store and retrieve documents in multiple languages. A secretary in a personnel department would pose a question, such as, “who has the experience with multilingual systems”. This question is looking for a series of potential answers. Information Retrieval System must be able to receive different kinds of queries and answers in different ways, retrieving individual facts or groups of potentially relevant items.

Such flexibility is necessary because of the widely varying sources and reasons for needing information, which are rooted in what Belkin, Brooke and Oddy [1972] refer to as Anomalous State of Knowledge (ASK). These ASKs may represent the lack of a particular fact in the information seeker’s knowledge base, they may represent a much larger area of missing information, or they may represent a lack of knowledge structure. A small need can often be answered with a single fact, while a much larger or more unstructured area within one’s knowledge base may potentially be answered in a variety of ways by a number of facts or documents. Information base, whether databases or collections, must be structured and organised to meet the potential needs of the people who will use them, so that the means of information retrieval must also be selected or created with an understanding of what information needs will have to be met, and how people are likely to understand and use the system.

18.2.3 Structured Query Language (SQL)

Structured Query Language (SQL) is a query language used for accessing and modifying information in a database. Some common SQL commands include ‘insert,’ ‘update,’ and ‘delete’. Queries take the form of a command language that lets a user select, insert, update, find out the location of data, and so forth. There is also a programming interface. The language was first created by IBM in 1975 and was called SEQUEL for “Structured English Query Language”. Since then, it has undergone a number of changes, with a lot of influence from Oracle Corporation. Today, SQL is commonly used for Web database development and management. Though SQL is now considered to be a standard language, there are still a number of variations of it, such as mSQL and MySQL. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language. Some Information Retrieval Systems are limited to finding those facts or documents containing characteristics specified by the query. Such systems are often referred to as database systems or database as SQL or forms variant of it recognized by the system under consideration. SQL allows precise specification of the value for attributes of terms to be retrieved.

An example of an SQL query might be:

Select from courses where Student’s Name = “Anjali Kapoor” and Department = “Management Science”

Self Check Exercise

1) What do you understand by Information Retrieval?

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

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

.....

.....

.....

.....

.....

18.3 INFORMATION RETRIEVAL TECHNIQUES

The organisation of information and the development of various techniques to retrieve information have been major areas of research. With the development in computer technology, interest in this area has been renewed through greater emphasis on the computerised information retrieval systems.

The required techniques can be broadly categorised into exact match, best match and partial match techniques and the latter can be divided into individual or network techniques. These can be further broken down to accommodate specific techniques, such as, cluster, probabilistic, vector-space and so on. The technique most widely used is the **exact match retrieval** one. It is implemented as Boolean, full-text or string searching. Its advantages are that it is easy to understand and use and is available on most systems. But its disadvantages are many. It misses many relevant texts, which match the query only partially; it neither ranks documents nor does it take into account the relative concepts either within the query or within the text. We may have to think about a researcher while looking for leads or at the beginning of the research problem, be satisfied with information retrieved using only this technique. In research of multi-dimensional and multi-disciplinary nature, and sometimes covering fringe areas, probably a **spreading activation technique** or the use of a citation strategy where the seed document is a highly cited one, will be more effective in this instance. The **best match** refers to comparison of two or more and need not be same as exact match. You may refer to Unit 17 for further details.

The issue of matching queries to retrieval has another aspect of study. The issue here is whether **specific retrieval techniques** for specific kinds of queries are on a professional basis of appropriateness and is preferable, or we can also opt for a single retrieval technique to many questions.

However, in the context of intelligent information systems, different types of retrieval techniques would be required for different situations of users. Cognitive mapping and other retrieval techniques provide an integrative approach to overall research process.

Self Check Exercise

2) What are the different Information Retrieval techniques?

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

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

.....

.....

.....

.....

.....

18.4 MODELS BASED ON INPUT/OUTPUT

On the basis of input and the output, Information Retrieval Models can be grouped into three basic categories:

- i) Data Retrieval Model
- ii) Information Retrieval Model
- iii) Knowledge Retrieval Model.

18.4.1 Data Retrieval Model

Data retrieval model essentially handles data. For the purpose of our understanding, data can be taken as unprocessed information or preliminary phase of information. Data is an unbiased fact which can be used to form an information. For example, we can say that the population of the city of Jaipur is eighty (80) lakhs. This is a data. Thus, a census system is a data retrieval system. Similarly, National Sample Survey Organization and Central Statistical Organisation can be taken to be numerical data systems. A data retrieval model calls for organisational structure based on various criteria such as properties, clusters and other different entities. There is a need for a taxonomic presentation of these aspects. Such a taxonomic presentation must also be accessible from other types of associations. A searcher of a data comes for a specific information retrieval. Therefore, the expression of information need should be very precise. Therefore, the data retrieval model is a simple model of information retrieval needing specific matching techniques viz., a taxonomic structure of the various entities involved and their properties.

18.4.2 Information Retrieval Model

Information is data oriented to a purpose. It actually combines several data into a relational structure. Information retrieval is, therefore, a more complex model. It has to generally comprehend multi-dimensional relationships. It is not amenable easily to a taxonomic structure. The representation of information is to be based on a relational data base structure using some associative mathematics. The expression of information need is also complex and time consuming. It draws out for a long conversational or browsing process and the information retrieval model must incorporate such facilities and interfaces.

18.4.3 Knowledge Retrieval Model

Knowledge is a kind of integration of general types of information. It normally occurs in the human mind. The human mind infers and integrates several coordinates with the information received by it. So, knowledge is assimilated information. In order to facilitate decision-making and problem solving, intelligent knowledge based information retrieval models are coming up. Such systems comprise three basic aspects:

- a) The so-called **knowledge base** or a store of accumulated set of rules for converting information into knowledge. It also incorporates knowledge acquisition system.
- b) The second aspect of the system is **inference engine**. An inference engine is capable of deriving appropriate information from the combination of rules for deriving a synthesized knowledge. This process of deriving is based on inferential logic using quantitative and non-quantitative techniques.
- c) A **user interface**, i.e., conversational process in the model which is capable of receiving information in the conversation mode and converting it into database signals for interaction purposes. Thus, a knowledge retrieval model is a sophisticated model of information processing, organization and retrieval.

Self Check Exercise

- 3) Discuss briefly the various Retrieval Models.

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

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

.....

.....

.....

.....

.....

18.5 MODELS BASED ON THEORIES AND TOOLS

Based on theories and methods/tools available in other disciplines, a number of models have been developed in order to find satisfactory solutions for information retrieval problems.

18.5.1 Boolean Retrieval Model

18.5.1.1 Standard Boolean Retrieval Model

Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. Typically, the Boolean functions apply to processing tokens identified anywhere within an item. The typical Boolean operators are AND, OR, and NOT. These operations are implemented using set intersection, set union and set difference procedures. A few systems introduced the concept of 'Exclusive OR' but it is equivalent to a slightly more complex query using the other operators and is not generally useful to users since most users do not understand it.

The normal Boolean operations produce the following results:

"A AND B" retrieves those items that contain both terms A and B.

"A OR B" retrieves those items that contain the term A or the term B or both.

"A NOT B" retrieves those items that contain term A and not contain term B.

18.5.1.2 Weighted Boolean Retrieval Model

The two major approaches to generating queries are Boolean and natural language. Natural language queries are easily represented within statistical model and are usable by the similarity measures. Issues arise when Boolean queries are associated with weighted index systems. Some of the issues are associated with how the logic (AND, OR, NOT) operators function with weighted values and how weights are associated with the query terms. If the operators are interpreted in their normal interpretation, they act too restrictive or too general. Salton [1979] showed that using the strict definition of the operators would sub-optimize the retrieval expected by the user. Closely related to the strict definition problem is the ranking which is missing in pure Boolean process. Salton provided additional insight into the issues of merging the Boolean queries and weighted query terms under the assumption that there are no weights available in the indexes. The objective is to perform the normal Boolean operations and then refine the results using weighting techniques.

Weighting of index terms is not common in manual indexing systems. Weighting is the process of assigning an importance to an index term's use in an item. The weight should represent the degree to which the concept associated with the index term is represented in the item. The weight should help in discriminating the extent to which the concept is described in items of the database. The manual process of assigning weights adds additional overhead on the indexer and requires a more complex data structure to store the weights. In a weighted indexing system, an attempt is made to place a value on the index term's representation of its associated concept in the document. An index term's weight is based upon a function associated with the frequency of occurrence of the term in the item. Typically, values for the index terms are normalised between zero and one. The higher the weight, the more the term represents a concept discussed in the item. The weight can be adjusted to account for other information such as the number of items in the database that contain the same concept.

The query process uses the weights along with any weights assigned to terms in the query to determine a scalar value (rank value) used in predicting the likelihood that an item satisfies the query. The results are presented to the user in order of the rank value from highest number to lowest number.

If weights are assigned to the terms between the values 0.0 to 1.0, they may be interpreted as the significance that users are placing on each term. The value 1.0 is assumed to be the strict interpretation of a Boolean query. The value 0.0 is interpreted to mean that the user places little value on the term. Under these assumptions, a term assigned a value of 0.0 should have no effect on the retrieved set. Thus,

“A1 OR B0” should return the set of items that contain A as a term.

“A1 AND B0” will also return the set of items that contain term A.

“A1 NOT B0” also return set A.

This suggests that as the weight for term B goes from 0.0 to 1.0 the resultant set changes from the set of all items that contains term A to the set normally generated from the Boolean operation. The process can be visualised by use of the Venn diagram shown in Figure 18.1.

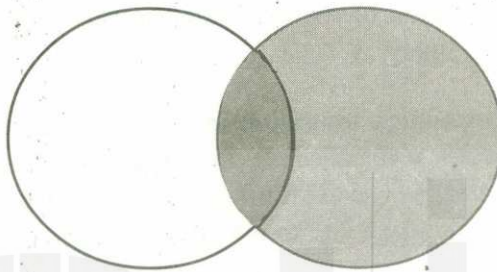


Fig. 18.1: Venn Diagram

Under the strict interpretation “A1 OR B1” would include all items that are in all the areas in the Venn diagram. “A1 OR B0” would be only those items in A (i.e., the white and dark shaded areas) which is everything except items in “B NOT A” (the grey area). Thus, as the value of query term B goes from 0.0 to 1.0, items from “B NOT A” are proportionally added until at 1.0 all of the items will be added.

Similarly, under the strict interpretation “A1 AND B1” would include all of the items that are in the black dotted area. “A1 AND B0” will be all of the items in A as described above. Thus, as the value of query term B goes from 1.0 to 0.0 items will be proportionally added from “A NOT B” (white area) until at 0.0 all of the items will be added.

Finally, the strict interpretation of “A1 NOT B1” is white area while “A1 NOT B0” is all of A. Thus as the value of B goes from 0.0 to 1.0, items are proportionally added from “A AND B” (dark shaded area) until at 1.0 all of the items have been added.

The final issue here is the determination of which items are to be added or dropped in interpreting the weighted values.

18.5.2 Fuzzy Logic Model

An Information Retrieval System has software component that has the features and functions required to manipulate ‘information’ items versus a DBMS that is optimized to handle ‘structured’ data. Here information is regarded as fuzzy text. The term ‘fuzzy’ is used to imply the results from the minimal standards or controls on the creators of the text items. The author presents concepts, ideas and abstractions along with supporting facts. As such, there is minimal consistency in the vocabulary and styles of items. The searcher has to be omniscient to specify all search term possibilities in the query.

Fuzzy logic supports values – true and false as well as other values in between.

The conceptual fuzzy logic was introduced by Professor Lotfi A Zadeh. The basic objective of the fuzzy logic is to develop a model that could be close to natural language process. It is an appropriate tool for modeling the kind of uncertainty associated with vagueness with imprecision.

Fuzzy retrieval provide the capability to locate spellings of words that are similar to the entered search term. This function is primarily used to compensate for errors in spelling of words. Fuzzy retrieval increases recall at the expense of decreasing precision. In the process of expanding a query term, fuzzy retrieval includes other terms that have similar spellings, giving more weight to words in the database that have similar word lengths and position of the characters as the entered term. A fuzzy search on the term 'computer' would automatically include the following words from the information database: 'computer', 'compiter', 'computter', 'compute'. An additional enhancement may lookup the proposed alternative spelling and if it is a valid word with a different meaning, include it in the search with a low ranking or not include it at all (e.g., 'commuter'). Systems allow the specification of the maximum number of few terms that the expansion includes in the query.

Fuzzy retrieval has its maximum utilisation in a system that accepts items that have been optical character recognised. In the OCR process, a hardcopy item is scanned into a binary image. The OCR process is a pattern recognition process that segments the scanned in image into meaningful sub-regions, often considering a segment – the area defining a single character. The OCR process will then determine the character and translate into an internal computer encoding. Based upon the original quality of the hardcopy this process introduces errors in recognising characters. With decent quality input, systems achieve in the 90-99 per cent range of accuracy. Since these are character errors throughout the text, fuzzy retrieval allows location of items of interest compensating for the erroneous characters.

18.5.3 Set Theoretic Model

The set theoretical view of information retrieval is based on the recognition that information requests are normally formulated by choosing collections or sets of item identifiers, or keywords. The keyword sets 'in turn' lead to the retrieval of record subsets chosen from among the stored collection of records. The fundamental data of retrieval theory are provided in this view by the relations which exist between the set of item descriptions and the corresponding record sets.

18.5.4 Vector Space Model

Often weighted systems are discussed as vectorised information systems. This association comes from the SMART system at Cornell University, created by Dr. Gerald Salton [1979]. The system emphasises weights as a foundation for information detection and stores these weights in a vector form. In systems, based upon a vector model, the semantics of every item are represented as a vector. A vector is a one-dimensional set of values, where the order/position of each value in the set is fixed and represents a particular domain. Each vector represents a document and each position in a vector represents a different unique word (processing token) in the database. There are two approaches to the domain of values in the vector – binary and weighted. Under the binary approach, the domain contains the value of one or zero, with one representing the existence of the processing token in the item. In the weighted approach, the domain is typically the set of all real positive numbers. The value of each processing token represents the relative importance of that processing token in

representing the semantics of the item. The value assigned to each position is the weight of that term in the document. A value of zero indicates that the word is not in the document. The system and its associated research results have been evolving for over 30 years. Queries can be translated into the vector form. Search is accomplished by calculating the distance between the query vector and the document vector. The use of weights also provides a basis for determining the rank of an item. The vector approach allows for a mathematical and a physical representation using a vector space model.

In addition to the general problem of dynamically changing databases and the effect on weighting factors, there are problems with the vector model on assignment of a weight for a particular processing token to an item. A major problem comes in the Vector Space Model when there are multiple topics being discussed in a particular item. There is no way to associate correlation factors between terms, since each dimension in a vector is independent of the other dimensions.

The Vector Space Model procedure can be divided into three stages. The first stage is the document indexing where the content bearing terms are extracted from the document text. It is obvious that many of the words in a document do not describe the content, like, the, is, are, in, to, of, etc. These are called non-significant words or stop words. In case of automatic indexing, these terms are removed from the document vector, so the document will only be represented by the content-bearing terms. In general, 40-50% of the total number of words, in a document, are stop words. These can be removed with the help of a stop word list. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

The VSM is contrary to the Boolean Retrieval Model in which retrieval is based on the hundred percent (exact) match. The VSM allows retrieval of the most similar to the query without the exact match. Thus, the VSM can be well explained in terms of *keyword-by-document matrix (A)*, in which the rows correspond to keywords (W) in the database and the columns correspond to documents (D), then the matrix will be like:

	D1	D2	D3	D4		Dn
W1	A11	A12	A13	A14		A1n
W2	A21	A22	A23	A24		A2n
A = W3	A31	A32	A33	A34		A3n
W4	A41	A42	A43	A44		A4n
.....						
Wm	Am1	Am2	Am3	Am4		Amn

Let us take a hypothetical example, like, an information seeker searches information on "Education information retrieval system". He uses four keywords W1, W2, W3, and W4. After searching the database, he gets six articles: A1, A2, A3, A4, A5, and A6. After analysis, it is found that the first article A1 talks only about W1; article A2 discusses one-third topic of W2 and two-third topic of W4; article A3 deals with 20% of W1, 30% of W3 and 50% of W4; article A4 deals with 60% of W1, 10% of W2 and 30% of W4; Article A5 talks 80% about W2 and 20% about W3; and article A6 discusses only about W4. Now this can be denoted in the form of a 4X6 matrix:

	A1	A2	A3	A4	A5	A6
W1	1.00	0.00	0.20	0.60	0.00	0.00
W2	0.00	0.33	0.00	0.10	0.80	0.00
A = W3	0.00	0.00	0.30	0.00	0.20	0.00
W4	0.00	0.67	0.50	0.30	0.00	1.00

The VSM is a retrieval model which constitutes a fairly large class of retrieval methods, each consisting of an indexing method and a retrieval function. The indexing method generates description vectors, and the retrieval function generates retrieval status values by comparing the query description vector with the document description vectors. A conceptual diagram of VSM is given at Figure 18.2. The information seeker is assumed to have information need, which he formulates as a query. The query q and the document d_j are indexed in two steps. First appropriate indexing features are spotted in the query q and in the document d_j . Secondly, these features are assigned weights to obtain the query description and the document descriptions are sets of weighted indexing features. These are called document description vector and query vector. The query description and document descriptions are matched and a score is generated for every document pair. These scores are called Retrieval Status Values (RSVs). For every query, the documents are presented to the information seeker in descending order of these RSVs.

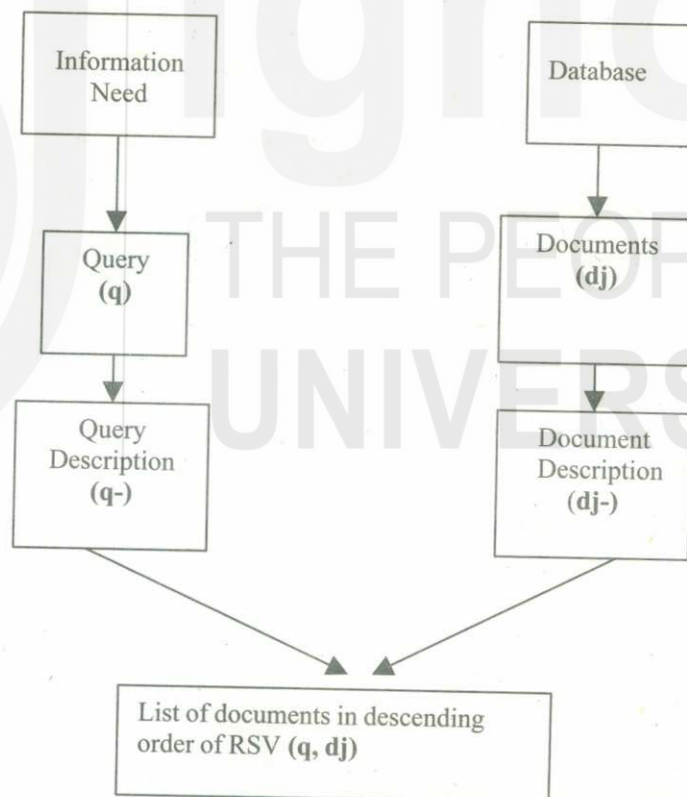


Fig. 18.2: Vector Space Model

The VSM relies on the premise that the meaning of a document can be derived from the document's constituent terms. Each keyword in a document collection forms document vector which represents the single or multiple occurrences of the term i in document d . Similarly, a query is represented by a query vector which denotes the number of occurrences of terms in the query. Both the document vector and query vector provide the locations of the objects in the term-document space. There are two common one-dimensional measures that every vector has, length and angle with respect to a fixed point. The angle

between two vectors refers to the measure in degrees between those two vectors. The document vector whose angle is closest to the query vector's angle is the best choice, yielding the document most closely related to the query. It is measured in terms of cosine angle between the two vectors. If the cosine of the angle is 1, then the angle between the document vector and the query vector measures 0 degree, meaning the document vector and the query vector move in the same direction. A cosine measure of 0 would mean the document is unrelated to the query vector. Thus, a cosine measure close to 1 means that the document is closely related to the query.

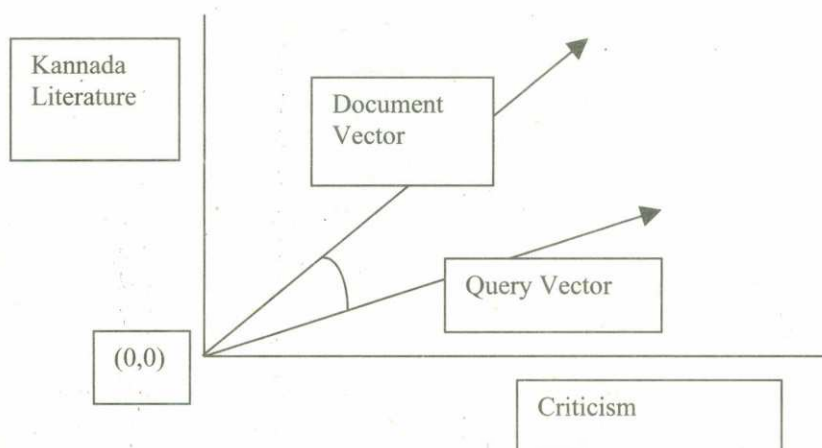


Fig. 18.3: Mapping of vector spaces

If a query (q) is considered to be a line in an imaginary space and the document (d) is also considered to be a line in the imaginary space, the geometrically determined angle between the two lines can be understood as measuring the degree to which the documents are similar to the query. While in the case of a large angle the document is presumed to be dissimilar to the query, in the case of a very small angle the document is presumed to be highly similar to the question. In this example, each Vector is plotted from (0,0) point to a point determined by the number of occurrences of the terms 'Kannada Literature' and 'Criticism' in the document or question. If there is one occurrence of 'Criticism' and two occurrences of 'Kannada Literature' in the document, the vector (d) in the Figure 18.3 is obtained.

18.5.5 Probabilistic Retrieval Model

Probabilistic considerations may apply if one assumes that system characteristics such as the term assignment to the records, or the relevance properties of the records are probabilistic in nature. Other mathematical techniques that have been used include decisions theory, information theory, pattern classification, mathematical linguistics and feature selection methods.

In addition to a vector space model, the other dominant approach uses a probabilistic retrieval model. The model that has been most successful in this area is the Bayesian approach. This approach is natural to information systems and is based upon the theories of evidential reasoning (drawing conclusions from evidence). Bayesian approaches have long been applied to information systems. The Bayesian approach could be applied as part of index term weighting, but usually is applied as part of the retrieval process by calculating the relationship between an item and a specific query.

The probabilistic approach is based upon direct application of the theory of probability to information retrieval systems. This has the advantage of being able to use the developed formal theory of probability to direct the algorithmic development. The use of probability theory is a natural choice because it is the basis of evidential reasoning.

According to Van Rijsbergen [1979], probability theory is an intuitively pleasing model for describing and analysing information retrieval. In this approach one can estimate the probability of retrieval. In other words, the probability of relevance of retrieval is measured. Van Rijsbergen proposed that a measure of the probability of relevance of a given document to a particular theory be based on a Vector representing that document. He postulated that the pattern of index terms in relevant documents will differ from their pattern in non-relevant documents. Probability approach will help to analyse term-clustering, frequency weighting, relevance weighting and ranking.

Probability theory can also be used to rank, and order documents according to their probability of relevance. Robertson [1978] shows that the order of documents can be based on term values and on 'Optimal Retrieval Function'. However, if one attempts to rank order of the documents in a Boolean environment, some difficulties arise which are inherent to the Boolean logic. Bookstein [1978] suggested that the retrieved documents be ordered according to the number of Boolean expressions present in the document that are true.

Doszkoc [1978] suggests that probability associations are being used to find terms that are associated with other terms. The association procedures are based on term occurrences and the frequency of these terms in the database.

The probabilistic approach of a retrieval model is based on the assumption that the distribution of the indexing features tells something about the relevance of a document. This approach adopts a retrieval model which optimizes the retrieval effectiveness according to the Probability Ranking Principle (PRP).

William Cooper has formulated the Probability Ranking Principle as "If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data". In this method the system replies to a query by presenting the beginning of a list of documents that are ranked in descending order of scores that either represent probabilities themselves or could be mapped to probabilities by means of an order preserving transformation. These scores often called Retrieval Status Values (RSV) depend on document descriptions consisting of appropriate statistical information about the indexing features. Such score may also depend on domain dependent parameters that are estimated by means of additional data e.g. by a thesaurus.

Probabilities are usually based upon a binary condition – an item is relevant or not. But in information systems the relevance of an item is a continuous function from non-relevant to absolutely useful. The output ordering by rank of items based upon probabilities, even if accurately calculated, may not be as optimal as that defined by some domain specific heuristic. The source of the problems that arise in application of probability theory come from a lack of accurate data and simplifying assumptions that are applied to the mathematical model.

18.5.6 Linguistic Model

In linguistic model for information retrieval, study the information retrieval from the point of view of the properties of language. Information retrieval is provided by features of natural language as well as artificial language. The various ways of storage of information are essentially based on natural language. The human

communication itself is full of natural language. In short, the languages carry three types of functions:

- i) They represent the contents of documents and other forms of information;
- ii) The information problem of users are represented in terms of language; and
- iii) Language is used in computer processing and in searching and retrieving of information.

The language works on three bases:

- a) Semantic basis which conveys meaning from one human being to another;
- b) The syntactic basis which helps formation of semantics in the use of grammar; and
- c) The vocabulary, which supply different meaning to terms for formation of sentences, paragraphs and other structures.

The logical structure of a language and the taxonomy of the languages refers to relationship between vocabulary and concepts. The vocabulary generally refers to the logical structure. In modern times the vocabulary control also include thesaural control and technical glossary control. Use of transformational grammar as well as parsing techniques provide processing speed of the language for information retrieval. Besides this, indexing language with coordinative control provides a basic model for information retrieval. Use of associative mathematics in search logic and in search expression formulation, provide yet another type of language control in information retrieval. This linguistic model forms an essential base for information retrieval. In social science field, language plays an ambiguous role because the terminology of the field is not as rigorous as in the field of natural sciences.

18.5.7 Mathematical Model

Mathematical model generally pre-supposes a careful formal analysis of the problem and specification of the assumptions and explicit formulation of the way in which the model depends on the assumptions.

Mathematical models are essentially based on representative mathematics as well as associative connections. In particular, cluster analysis and clustering techniques are used on experimental basis in automatic abstracting and indexing. Use of sets theory and Boolean logic is a very familiar method of mathematical modeling of information. Concept of similarity measures and choice of variable and the combinational aspects of clustering try to provide semantic structure for information represented. Cluster analysis today involves statistical packages or clustering software.

18.5.8 Psychological Model

The psycholinguistic approaches to information retrieval led to the study of formation of concepts in human mind the way in which the human thinking process arranges the ideas, its presentation at the time of enquiry, and the type of retrieval cues it demands while searching has led to a cognitive research linked with computer communication processes. The studies of Belkin, Brooks and Oddy [1979] on Anamolous State of Knowledge provides interesting insights in relation to information retrieval process. Further, the current day studies in the field of information retrieval and artificial intelligence have thrown sufficient light to bring in harmonious coupling of psychological theory into information retrieval.

18.5.9 Economic Model

The economic model of information retrieval centres round the measures of cost effectiveness and cost efficiency of information retrieval. These two criteria are based on performance of information retrieval systems in relation to input cost as well as the number of successful outputs. The concept of provision of multiple access points being used gives a chance for measurement of information transfer. The field of information retrieval, which has developed several models of information measurement based on statistical and mathematical techniques used for studies in bibliometrics and scientometrics provides a scope for correlation of economic benefits. However, due to various intangible elements in information retrieval, which cannot be identified, the economic model does not yet provide a holistic approach to information retrieval.

Theoretically, the modelling of information retrieval can be looked at from output such as data, from the methodological approaches from different disciplines. But, in practice we may not find a single model operating in that manner. For that purpose, a collective modelling of various levels can be seen through. For example, the ERIC (Educational Resources Information Center) model for IR is one of the important areas of IR, which is a combination of all approaches to information retrieval [Henry and Diodate, 1991]. The idea behind these theoretical models is to help analytical studies for bringing in efficiency to different aspects of information retrieval. These multi-disciplinary approaches to information retrieval provide a better base.

18.5.10 Hypertext Linkage Model

Hypertext linkages are creating an additional information retrieval dimension. Traditional items can be viewed as two dimensional constructs. The text of the items is one dimension representing the information in the items. Imbedded references are a logical second dimension that has had minimal use in information search techniques. The major use of the citations has been in trying to determine the concepts within an item and clustering items. Hypertext, with its linkages to additional electronic items, can be viewed as networking between items that extend the contents. The imbedding of the linkage allows the user to go immediately to the linked item for additional information. The issue is how to use this additional dimension to locate relevant information.

Looking at the Internet at the current time there are three classes of mechanisms to help find information: manually generated indexes or directories, automatically generated indexes and web crawlers (intelligent agents). Yahoo (<http://www.yahoo.com>) is an example of the first case where information sources (home pages) are indexed manually into a hyperlinked hierarchy. The user can navigate through the hierarchy by expanding the hyperlink on a particular topic to see the more detailed sub-topics. At some point the user starts to see the end items. Lycos (<http://www.lycos.com>) and AltaVista (<http://www.altavista.com>) automatically go out to other Internet sites and return the text at the sites for automatic indexing. Lycos returns home pages from each site for automatic indexing while AltaVista indexes all of the text at a site.

Web crawlers (WebCrawler, Open Text, Pathfinder) and intelligent agents (Coriolis Groups' NetSeeker) are tools that allow a user to define items of interest and they automatically go to various sites on the Internet searching for the desired information. The Uniform Resource Locator (URL) hypertext links can map to another item or to a specific location within an item.

Self Check Exercise

- 4) Write down the different Information Retrieval Models based on theories and tools.

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

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

.....

.....

.....

.....

.....

18.6 SUMMARY

An information retrieval system (IRS) is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. It consists of a software program that facilitates a user in finding the information the user needs. An IRS includes database, information base and Structured Query Language (SQL). Database consists of a set of records or files. An information base consists of a set of databases. SQL is a standard interactive and programming language for getting information from and updating a database. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language. Queries take the form of a command language that lets user select, insert, update, find out the location of data, and so forth. Information retrieval systems process users queries as well as manipulate online databases using SQL. Thus, it is very important area of study. Retrieval models, like, Data Retrieval Model, Information Retrieval Model and Knowledge Retrieval Model are based on input and output. Data Retrieval Model handles data and can be taken as unprocessed information or preliminary phase of information. Census system is a data retrieval system. Information Retrieval Model is data oriented to a purpose. It combines several data into a relational structure. Knowledge Retrieval Model assimilates several types of information in order to facilitate decision-making and problem solving. It is a sophisticated model of information processing and organisation. The IR models based on theories and tools try to develop efficient IR systems.

18.7 ANSWERS TO SELF CHECK EXERCISES

- 1) Information Retrieval is a process of selecting information from a store.
- 2) Information retrieval techniques are broadly categorised into three types:
 - i) Exact match retrieval:
 - ii) Best match retrieval
 - iii) Partial match retrieval.
- 3) There are three basic information retrieval models.
 - i) **Data Retrieval Model** – It essentially handles data and can be taken as unprocessed information or preliminary phase of information. Census system is a data retrieval system.
 - ii) **Information Retrieval Model** – It is data oriented to a purpose. It combines several data into a relational structure.

- iii) **Knowledge Retrieval Model** – It assimilates several types of information in order to facilitate decision-making and problem solving. It is a sophisticated model of information processing and organisation.

4) The different information retrieval models based on theories and tools are:

- Boolean Retrieval Model (Standard and Weighted)
- Fuzzy Logic Model
- Set Theoretic Model
- Vector Space Model
- Probabilistic Retrieval Model
- Linguistic Model
- Mathematical Model
- Psychological Model
- Economic Model
- Hypertext Linkage Model

18.8 KEYWORDS

- Anomalous State of Knowledge (ASK)** : Lack of a particular fact in the information seeker's knowledge base, it may represent a much larger area of missing information, or it may represent a lack of knowledge structure.
- Fuzzy Retrieval** : The capability to locate spellings of words that are similar to the entered search term. Fuzzy retrieval increases recall and decreases precision. In the process of expanding a query term, fuzzy retrieval includes other terms that have similar spellings. This function is primarily used to compensate for errors in spelling of words.
- Retrieval Status Values (RSVs)** : In retrieval process, the query description and document descriptions are matched and a score is generated for every document pair. These scores are called Retrieval Status Values (RSVs).
- Weighting** : Weighting is the process of assigning an importance to an index term's use in an item. The weight represents the degree to which the concept associated with the index term is represented in the item. In a weighted indexing system, an attempt is made to place a value on the index term's representation of its associated concept in the document.

18.9 REFERENCES AND FURTHER READING

- Belkin, N.J., Brooks, H.M. and Oddy, R.N. (1979). *Representing and classifying Analogous States of Knowledge*. In: *The analysis of meaning: informatics*, edited by M. Maccafferty and K. Gray. London: Aslib.
- Bookstein, Abraham. (1978). On the perils of merging Boolean and weighted retrieval system. *Journal of American Society for Information Science*, 29(3), 156-8.
- Brookes, Bertram C. and Griffiths, Joseph. (1978). *Frequency rank distributions*. *Journal of the American Society for Information Science*, 27(1), 13-17.

- Carter, M.B. (1986). A methodology for the economic appraisal of management information. *International Journal of Information Management*, 193-203.
- Doszkoc, Tamas E. (1978). AID: an Associative Interactive Dictionary of online searching. *Online Review*, 2(2), 163-73.
- Gopinath, M.A. (1999). *Information retrieval*. In: MLIS-03 course materials. Unit 13. New Delhi: Indira Gandhi National Open University.
- Gopinath, M.A. (1999). *ISAR systems: operations and design*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.
- Gopinath, M.A. (1999). *Objectives of information storage and retrieval systems*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.
- Henry, G. and Diodato, V. (1991). The rates of assignment of thesaurus terms in the ERIC information retrieval system: an analysis of hierarchies and levels. *Journal of Documentation*, 47(3), 276-283.
- Karen, Spark Jones (1973). *Linguistics and information science*. New York: Academic.
- Kemp, Alister (1988). *Knowledge base retrieval system*. London: Aslib.
- Levitan, K.B. (1982). Information resources as 'Goods' in the life cycle of information production. *Journal of American Society of Information Science*, 33, 44-54.
- McGill, Michael J. (1978). Knowledge and information spaces: implications for retrieval systems. *Journal for the American Society for Information Science*, 27(4), 205-10.
- Mock, T.I. and Vasarhelyi, M.A. (1980). A synthesis of the information economics and lens model. *Journal of Accounting Research*, 477-505.
- Murthy, S.G.K. and Biswas, R.N. (2004). A fuzzy logic based search technique for digital libraries. *DESIDOC Bulletin of Information Technology*, 24(6), 3-10.
- Rijsbergen, Van C.J. (1979). *Probabilistic retrieval*. In: *Information retrieval*, edited by C.J. Van Rijsbergen. 2nd Ed. London: Butterworths.
- Robertson, Stephen E. (1978). On the nature of fuzz: a diatribe. *Journal of the American Society for Information Science*, 29(6), 304-7.
- Salton, Gerard. (1979). Mathematics and information retrieval. *Journal of Documentation*, 35 (1), 1-29.
- Salton, G. and McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Smith, L.C. (1976). Artificial Intelligence in Information Retrieval System. *Information Processing and Management*, 12(3) 189-222.
- Wikipedia.org. (2005). *Information retrieval*. <http://en.wikipedia.org/wiki/information_retrieval/>.

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:

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

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

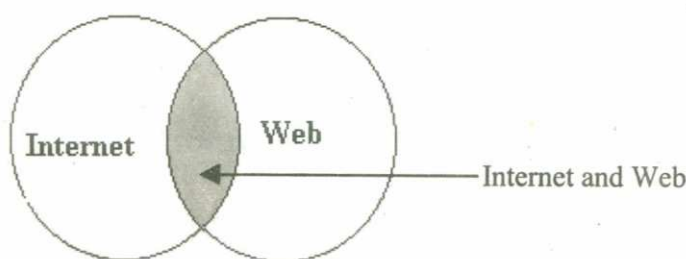
Many bibliographic information retrieval systems provide two types of search facilities for conducting an unknown item search: keyword search and subject search. A **keyword search** allows users to enter one or more keywords pertaining to their query. These keywords can be chosen by the user in any 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.

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



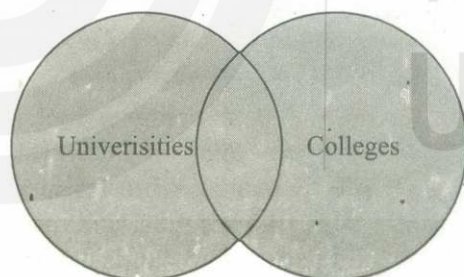
Search Items	Results	Remark
Internet	783	
Web	2962	
Internet AND Web	97	Shaded Area in the Figure

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

Internet and Web

Query: I would like information about colleges and universities.

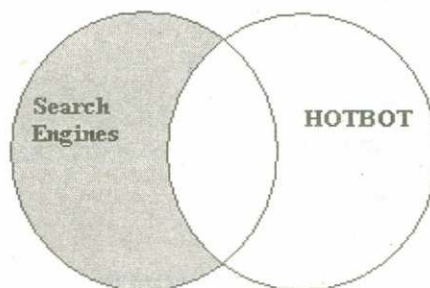


Search Items	Results	Remarks
Colleges	17320	
Universities	33685	
Colleges OR Universities	35702	Shaded Area in the Figure

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:

Limit Restriction	Qualifier	Example
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:

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

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

- **information providers** or database producers who provide databases to be accessed in an online mode;
- **a search service provider** or vendor, which provides access to the databases and software for conducting the search;
- **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
- **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:

- 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.
- 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.

- 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.
- 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:

- 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.
- Ingenta (<http://www.ingenta.com/>): Ingenta provides access to the full text of over 5,400 publications from over 230 academic and professional publishers.
- 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:

- company information – both directory listings and financial information
- 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
- patents and trademarks – a worldwide collection for research and competitive intelligence tracking
- chemistry, environment, science and technology – technical literature and reference material to support research needs
- social science and humanities including education, information science, psychology, sociology and science, from public opinion, news, and leading scholarly and popular publications
- general reference information – people, books, consumer news and travel.

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

- guided search mode, which does not require knowledge of the Dialog command language;
- command search mode, which allows experienced users to use the Dialog command language;
- database selection tools, which help users pinpoint the right database for a search;
- integrated database descriptions, pricing information and other search assistance; and
- 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:

- *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.
- *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:

- all of the databases together
- a group of similarly designed databases together
- 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:

- *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.
- *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:

- display specific records in more detailed formats or send records via e-mail or fax, or by post
- rearrange the order in which the records are displayed
- refine the search strategy
- remove duplicate records
- view the prices for all format options
- save the strategy for future use
- 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:

- a textbox for entering Dialog search commands
- a Submit button that sends the command
- 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:

- after log-in if it is set as the default
- when the Command Search link from the main Guided Search page is clicked
- 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.

Search operator (can be used lower or upper case)	Example	Description
Proximity operator (w)	S information (w) system S information (2w) system	Requests adjacent terms in the defined order. 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 S information (2n) system	Requests adjacent terms in any order. 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 code	S internet/ti,de	Suffix codes can only be used for the basic index fields
Prefix code	S au=Chowdhury, G.	Prefix codes are used to search the additional indexes
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 S EM>1000	In the first example the system will retrieve records with 100 employees; in the second with employees of more than 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.

- 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:

- 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.
- 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.

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:

- They search the Internet – or select parts of the Internet – according to a set of criteria.
- 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.
- 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.

Search features	Examples
Boolean search	Usually three different approaches are followed: ● combining search terms with AND, OR, NOT (e.g., ANDNOT in AltaVista) ● using '+' or '-' signs before the search terms ● 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:

- The 'I'm Feeling Lucky™' button in Google takes users directly to the first web page Google returned for a query.
- Site Collapse in AltaVista shows a maximum of two results per site, allowing the user to look at results from various sites.
- 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.
- TouchGraph GoogleBrowser: Shows users what the web 'looks like' to the search engine, displaying the links between the user's favourite websites.
- 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 crawls 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>>.



ignou
THE PEOPLE'S
UNIVERSITY



ignou

THE PEOPLE'S
UNIVERSITY