UNIT 17 INFORMATION RETRIEVAL PROCESSES AND TECHNIQUES

Structure

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.

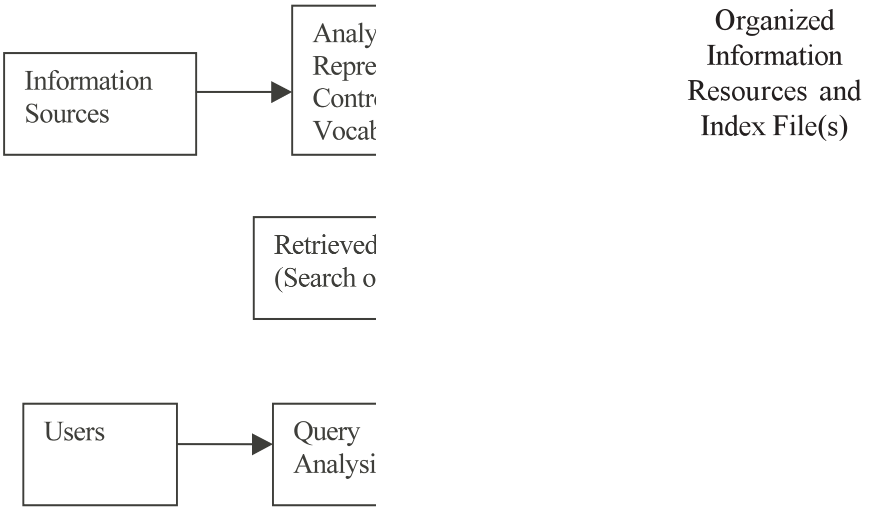

Fig. 17.1: Broad outline of an Information Retrieval System (IRS)

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	Adddison-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.

....

....

....

....

....

17.6 VOCABULARY CONTROL

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].

Table 17.2: The Four Eras of Debate on Controlled Vs. Natural Language Indexing

Era One –	controlled vocabulary
Era Two –	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.
Era Three –	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.
Era Four –	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.

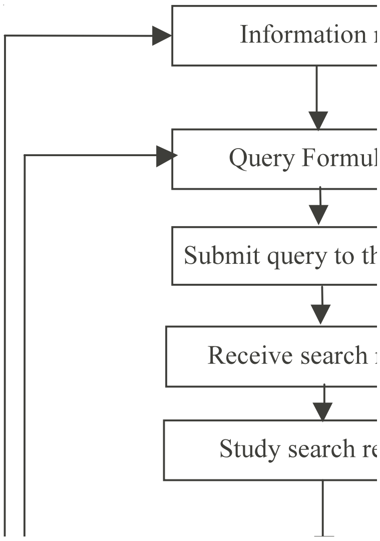

Fig. 17.4: Basic information access model

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.

17.9.4 Information Retrieval Features of e-Journal Services

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 *ChesireII* 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., Yahooligans, 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:

   ι to promote the consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials.

ι 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 Retrieval System : A system that carries out intelligent retrieval, i.e., 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.

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

Search Engine : 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.