15

INFORMATION RETRIEVAL SYSTEM: CONCEPT AND SCOPE

15.1 INTRODUCTION

Information is communicated or received knowledge concerning a particular fact or circumstance. Retrieval refers to searching through stored information to find information relevant to the task at hand. In view of this, information retrieval (IR) deals with the representation, storage, organization of/and access to information items. Here, types of information items include documents, Web pages, online catalogues, structured records, multimedia objects, etc. Chief goals of the IR are indexing text and searching for useful documents in a collection. Libraries were among the first institutions to adopt IR systems for retrieving information.

In this lesson, you will be introduced to the importance, definitions and objectives of information retrieval. You will also study in detail the concept of subject approach to information, process of information retrieval, and indexing languages.

15.2 OBJECTIVES

After studying this lesson, you will be able to:

- define information retrieval;
- understand the importance and need of information retrieval system;
- explain the concept of subject approach to information;

Notes

- illustrate the process of information retrieval; and
- differentiate between natural, free and controlled indexing languages.

15.3 INFORMATION RETRIEVAL (IR)

The term 'information retrieval' was coined by Calvin Mooers in 1950. It gained popularity in the research community from 1961 onwards, when computers were introduced for information handling. The term information retrieval was then used to mean retrieval of bibliographic information from stored document databases. But those information retrieval systems (IRS) were document retrieval systems. These were designed to retrieve information about the existence (or non-existence) of bibliographic documents relevant to a user's query. In other words, early IRS were designed to retrieve an entire document (a book, an article, etc.) in response to a search request. Although this is what today's IRS do, but over the years, many advanced techniques have been developed and applied to design the IRS. Over the years, the connotation of information retrieval has changed and it has been variously denoted by information professionals and researchers. Some of these include, information storage and retrieval, information organization and retrieval, information processing and retrieval, text retrieval, information representation and retrieval and information access.

Let us now understand the means through which information retrieval is carried out by libraries and some of the systems, for searching information from documents in its collection. No matter how large the collection, the library is of little value if it is unable to retrieve the right documents as and when required by a user. To do this, it must maintain an information retrieval system. When a match is achieved between the information requested and information in the retrieval system, then requested documents are located. In other words, the information supplied from the document(s) matches to an acceptable degree with the information demanded by the user. Achieving a successful match is the central objective of information retrieval.

The principal function of any library is to make available to the users, the information they need. In order to fulfill this function, the information which is stored in the library must be retrieved from the library database. Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Information retrieval is the process of selecting information from the stored information. The process is becoming increasingly dependent on computers and telecommunications technology. The design of information retrieval systems has presently become an important area of applied information technology.

INTEXT QUESTION 15.1

1. Why is Information retrieval an important function of any library?

15.4 INFORMATION RETRIEVAL SYSTEM

The concept of Information Retrieval System (IRS) is self-explanatory from the terminological point of view and refers to a 'system which retrieves information'. IRS is concerned with two basic aspects: (i) How to store information, and (ii) How to retrieve information.

One may simply denote such a system as one that stores and retrieves information. IRS is comprised of a set of interacting components, each of which is designed to serve a specific function for a specific purpose. All these components are interrelated to achieve a goal. The concept of IR thus is based on the fact that there are some items of information which have been organized in a suitable order for easy retrieval.

An information retrieval system is designed to analyze, process and store sources of information and retrieve those that match a particular user's requirements. 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 documents. IRS originally meant text retrieval systems as 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. Thus, modern information retrieval systems deal with storage, organization and access to text, as well as multimedia information resources.

Thus, an IR system is a set of rules and procedures, for performing some or all of the following operations:

a) Indexing (or constructing of representations of documents);

b) Search formulation (or constructing of representations of information needs);

c) Searching (or matching representations of documents against representations of needs); and

d) Index language construction (or generation of rules of representation)

So information retrieval is collectively defined as a "science of search" or a process, method and procedure used to select or recall, recorded and/or indexed information from files of data.

Notes

15.4.1 Objectives and Functions of IRS

The major objective of an IRS is to retrieve the required information whenever needed. It is 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.

The major functions of an IRS are:

(i) To identify the sources of information relevant to the areas of interest of the target users' community;

(ii) To analyze the contents of the sources (documents);

(iii) To represent the contents of the analyzed sources for matching with the users' queries;

(iv) To match the search statement with the stored database;

(v) To retrieve the information that is relevant; and

(vi) To make necessary adjustments in the system based on feedback from the users.

INTEXT QUESTION 15.2

1. What is the major objective of Information Retrieval System (IRS)?

15.5 IMPORTANCE OF INFORMATION RETRIEVAL

Libraries contain information in various physical forms. While for many users, the book is still a major vehicle for communication of information; for others, the periodical or the technical report have taken its place; and for yet others, films or gramophone records are significant. It is clear that the same work can appear in various physical forms. The intellectual content will be the same in each case, but obviously it is not practical to try to arrange the different physical forms together. We cannot, therefore, rely on the physical arrangement of the items in a library to gather different versions of the same work. We have to rely on a substitute – a set of records (surrogates) of the content of the library. These are in the form of library catalogues and bibliographies.

The library catalogue, however, is only one of the tools which serves as the key to library documents. A library containing a large number of periodicals will not attempt to list all articles of every issue if receives. Instead, we rely on indexes, abstracts and similar bibliographic tools which present the contents of periodicals as well. This enables us to obtain access to any particular item through number of approaches.

INTEXT QUESTION 15.3

1. Explain the importance of catalogues and bibliographic tools in libraries.

15.6 SUBJECT APPROACH TO INFORMATION

Users often approach information sources with a query that requires an answer or they seek information or documents on specific topics. This method of seeking information from sources by the users is referred to as subject approach to information. In order to provide this kind of information, it is necessary for information organizations to arrange documents or surrogates of documents in library catalogues, indexes or databases in such a way that items of specific information can be retrieved. There are various methods of providing information contained in documents using the subject approach. Two chief methods for the same are:

- Alphabetical subject approach
- Display of subject relationships

15.6.1 Alphabetical Subject Approach

Here the items of information are first grouped under the subject and then arranged according to alphabetic order so that specific subjects can be retrieved easily. Some problems to be overcome here are those related to synonyms, homographs, singular or plural forms, complex and compound words or subjects and multiword concepts.

15.6.2 Display of subject relationships

Like human beings, the subjects too have relationships, these include syntactic relationships and semantic relationships. Syntactic relationships deal with the way words and phrases of a sentences are arranged to show how they relate to each other. For example, a keyword search for “Photographs and Albums”, should allow users to specify whether they want “Photographs of Albums” or “Albums of Photographs”. Semantic relationships deals with the meanings of the words. For example, there is semantic difference between mercury (Planet) and mercury (metal), though two words are identical in sound and spelling.

The first librarian to consider the detailed arrangement by subject was Melvil Dewey. Librarians prior to Dewey had certainly arranged their libraries in classified order; the classified catalogue was well known. However, these classified arrangements were in broad subject groups; there was no attempt to give the detailed subject specification that Dewey suggested and which was necessary and useful. Dewey's classification scheme served two purposes: the first of these was the arrangement of books on shelves; and the second was the arrangement of entries in catalogues and bibliographies.

The Subject approach or subject indexing is the process or technique of identifying and selecting terms (words, phrases, sentences, taxonomic categories, notation) to indicate what a document is all about. It helps to summarize its contents and increases its retrieval by users. In other words, it is about identifying and describing the subject of documents. Its purpose is to facilitate finding a particular information on the basis of its subject content.

The two steps of subject indexing are:

a) Subject analysis to generate concepts that describe the document, and

b) Translation of concepts into controlled vocabulary for retrieval

INTEXT QUESTION 15.4

1. Describe how the subject approach to information came into existence.

15.7 INDEXING LANGUAGES

As discussed above, when the librarians apply subject approach to information, they are confronted with the difficult task of subject indexing. They have to deal with the complexity, variability, and richness of natural language of documents. Using unlimited or uncontrolled set of words or phrases to index leads to wasted efforts. There is also a high degree of searching failure due to vast range of words chosen by users. It is rightly said that no two words in a language mean exactly the same and there are no true synonyms. But words are often very close in meaning and more often not clearly understood. The inconsistent/varying words could lead to failure in searching as the users may not choose the words or terms that might be used by the indexer or the authors of the documents. In order to overcome various complex indexing problems, many forms of controlled vocabularies have been developed.

Retrieval of information by subjects from huge mass of documents requires that essential concepts are identified and organised in a searchable form. Indexing is a mechanism by which information contained in documents can be organised. But the problems lie with identifying and organising the concepts. In the documentary information, authors communicate in natural languages which are characterized by linguistic features. To overcome the problems of natural language, the need for an artificial language or indexing languages arises. It means that an indexing language is a language used for subject classification or indexing of documents. An Indexing language is defined as the set of terms used in an index to represent topics or features of documents, and the rules for combining or using those terms.

The purpose of an indexing language is to express the concepts of documents in an artificial language so that users are able to get the required information. The indexing language does this by depicting the relationships among the differently related concepts.

There are three main types of indexing languages.

1. Natural indexing language - Any term from the document in question can be used to describe the document.

2. Free indexing language - Any term (not only from the document) can be used to describe the document.

3. Controlled indexing language - Only approved terms can be used by the indexer to describe the document.

In the following sections, you will be introduced, in brief, to natural, free and controlled indexing languages.

15.7.1 Natural Indexing Language

Natural language refer to our language, which we normally use for communication. Whereas, languages that we design for a specific purpose or use in a specific sense or only for limited use are artificial languages. Natural indexing languages are thus 'natural language' or ordinary language of the document being indexed. Any term that appears in the document is a candidate for index terms. In practice, natural language indexing tends to rely upon the terms present in an abstract or the title of a document. Natural language indexing is based upon the full text of a document, depending on how it is archived. It may lead to very extensive indexing of each document or will involve establishing some mechanism for deciding which terms are the most important in relation to a particular document. In computerized indexing this will involve statistical analysis of the relative frequency of occurrence of terms. In human indexing some judgment would be required in selecting the terms. Many of these problems can be minimized by restricting indexing to titles and abstracts. Either, a computer or a person can execute natural language indexing. In computer indexing the computer may well use a list of terms deemed to be useful in indexing (example, a type of thesaurus) to identify appropriate terms. The use of natural language is depicted in a Figure 15.1

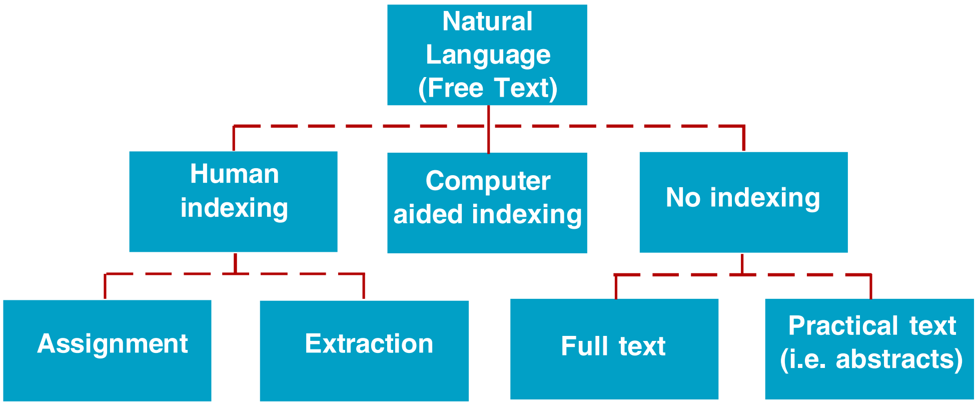

Fig 15.1 Natural Language Indexing

Any term that appears in the title, abstract or text of the 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. It is human language in which the structure and rules have evolved from usage, usually over a period of time. In search software designed to handle input expressed in natural language, the user may enter the query in the same form in which it would be spoken or written. Any term from the document in question can be used to describe the document.

15.7.2 Free Indexing Language

In free indexing language any term, not only from the document, can be used to describe the contents of a documents. Indexing is 'free' in the sense that there are no constraints on the terms that can be used in the indexing process. Free language indexing is distinct from natural language indexing in that natural language indexing is constrained by the language of the document being indexed; free language indexing does not even recognize these constraints. Free language indexing may be conducted by humans or computers. When executed by humans with a sound knowledge of a subject and its terminology, free language indexing can result in an index which is both consistent in the assignment of index terms, and which matches the perspective of users.

Notes

However successful, free language indexing is very dependent upon the skills of an individual indexer. Computerized free language indexing is, for all practical purposes, the same as natural language indexing.

It is the nature of a free indexing language that any word or term that suits the subject may be assigned as an indexing term. The terms may be machine or human assigned although free language is most common in a machine indexing environment. The computer operates by indexing every word with which it is provided unless it is instructed to do otherwise.

Controlled vocabularies usually improve the accuracy of free text searching, reduce irrelevant items in the retrieval list. Both natural language indexing and to some extent free language indexing are used in producing both printed indexes, computerized databases and databanks.

15.7.3 Controlled Indexing Language

Controlled indexing languages are indexing languages in which the terms used to represent subjects and the process by which terms are assigned to a particular document are controlled or executed by a person. Normally, there is a list of terms which acts as the authority list in identifying the terms that may be assigned to documents. An indexing involves a person assigning terms from this list to specific documents.

There are two types of controlled indexing languages: alphabetical indexing languages and classification schemes. In alphabetical indexing languages, such as, the thesauri and subject headings lists, subject terms are the alphabetical names of the subjects. Control is exercised over which terms are used, but otherwise the terms are ordinary words. In classification schemes, each subject is assigned a piece of notation. The usual objective of assigning notation is to place a subject within a context with respect to other subjects. Both classification schemes and alphabetical indexing languages are used in a variety of contexts. These devices are used in catalogues, indexes to books and periodicals, bibliographies, current awareness bulletins, selective dissemination of information, computerized databases, and databanks, abstract and indexing services, encyclopedias, dictionaries and directories. Classification is also prominent in the physical arrangement of documents.

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. An 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. So, it is only approved terms that can be used by the indexer to describe the document.

Notes

Compared to free text searching, the use of a controlled vocabulary increases the performance of an information retrieval system.

INTEXT QUESTION 15.5

1. Explain the term 'natural indexing language'.
2. What is the nature of a free indexing language?
3. Describe a Controlled indexing language.

WHAT YOU HAVE LEARNT

- The term 'information retrieval' was coined by Kelvin Mooers in 1950.
- Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.
- The term information retrieval was earlier used to mean retrieval of bibliographic information from stored document databases.
- Information storage and retrieval, information organization and retrieval, information processing and retrieval, text retrieval, information representation and retrieval and information access are different connotation of information retrieval.
- A library fulfills its function of information retrieval by maintaining some system for searching information out of documents from its collection.
- Modern information retrieval systems deal with storage, organization and access to text, as well as multimedia information resources.
- The major objective of an IRS 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.
- The first librarian to consider the detailed arrangement by subject was Melvil Dewey.
- Natural indexing language is not really a separate language but the 'natural language' or 'ordinary language' of the document being indexed.
- Free indexing language is not a listed language of terms, but the terms are provided by the indexer suitable to describe the contents of a document.

- Controlled indexing language is an indexing language in which the terms used to represent subjects, and the process whereby terms are assigned to a particular document, are controlled or executed by a person.

TERMINAL QUESTIONS

1. What are the objectives of an Information retrieval system?
2. Discuss the major functions of an Information retrieval system.
3. Distinguish natural, free and controlled indexing languages.
4. Explain the use of natural language with help of a diagram.

ANSWER TO INTEXT QUESTIONS

15.1

1. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. It is one of the most important functions of a library, because it meets the demands of required information of a user.

15.2

1. The major objective of an information retrieval system, is to retrieve the information. It is, either the actual information or through the documents containing the information surrogates that fully or partially match the user's query.

15.3

1. The library catalogue is a tool which indicates the availability and location of library documents. Catalogue does not provide information contained in the documents like articles in a periodical, etc. This information is provided by indexes, bibliographic abstracts and similar bibliographic tools in the library.

15.4

1. The first librarian to consider detailed arrangement by subject was Melvil Dewey. Librarians prior to Dewey had certainly arranged their libraries in classified order; the classified catalogue was well known. However, these classified arrangements were in broad subject groups; there was no attempt to give the detailed subject specification that Dewey suggested was necessary and useful.

15.5

1. In natural indexing language, the terms are selected from the same document to describe its content.

2. The nature of a free indexing language is that any word or term that suits the subject may be assigned as an indexing term.

3. Controlled indexing language is an indexing language in which the terms used to represent subjects and the process whereby terms are assigned to particular documents, are controlled or executed by a person.

GLOSSARY

Data Retrieval: The retrieval of information whose contents satisfy the information needs of user as per a user query.

Index Term: A pre-selected term which can be used to refer to the contents of a document.

Information Retrieval (IR): To find material (usually documents) that satisfies an information need from within large collections (usually stored on computers).

Keyword: Same as **Index Term**

Query: The expression of the user information need.

Retrieval: The task executed by an information system in response to a user request.

User Information Need: A natural language statement of an informational need of a user.

Vocabulary: Set of all the words in a text

WEBSITES

http://en.wikipedia.org/wiki/Information_retrieval

http://polaris.gseis.ucla.edu/pagre/is277.html

http://nlp.stanford.edu/IR-book/