

## **LESSON 6: AUTOMATING INFORMATION RETRIEVAL SYSTEM**

### **Structure:**

- 6.0. Objectives
- 6.1. Introduction
- 6.2. Development of automated Information Retrieval System
- 6.3. Purpose of Information Retrieval.
- 6.4. Advantages of automated information retrieval system
- 6.5. Functions of automated information retrieval system.
  - 6.5.1. Input
    - 6.5.1.1. Users
    - 6.5.1.2. Documents
  - 6.5.2. I.R.Procedures
    - 6.5.2.1. Automatic text analysis
    - 6.5.2.2. File structure
    - 6.5.2.3. Indexing language
    - 6.5.2.4. Search strategy
  - 6.5.3. Services of computerized Information Retrieval system
    - 6.5.3.1. Reference service
    - 6.5.3.2. Retrospective services
    - 6.5.3.3. CAS / SDI
    - 6.5.3.4. Internet – based Information Retrieval
    - 6.5.3.5. Support services
- 6.6. Feedback
- 6.7. Evaluation of information retrieval system
- 6.8. Summary
- 6.9. Further Readings
- 6.10. Unit-end-questions
- 6.11. Answers to unit-end-questions.

### **6.0. Objectives**

In this unit you will be introduced to the concept of automate information retrieval activities and services. After reading this unit, you will be able to

- Explain the origins of computerized information retrieval;
- Understand the functions of a computerized information retrieval system;
- Identify various services that can be produced and delivered;
- Benefits to be reaped from automation of information retrieval system.

### **6.1. Introduction**

Information retrieval has always been the focus of information scientists and reference librarians, trying to provide relevant materials in response to a user's query. As reference librarians, we often pride ourselves on knowing our collections and what it holds. The purpose of information retrieval is to provide useful information in response to a query. Despite the methods used, whether it is manual search to locate the information, or the computer searches its index, it can be achieved through a variety of products and services. Computerized information retrieval system exploits information content in all types of documents to provide a variety of services to user community. Often information is also extracted from external sources besides in-house database or OPAC.

Computers can be applied in two ways in information retrieval for providing various information services. Firstly, it can be used to build up an in house information storage and retrieval system, which is limited to the resources available in the institution to give information services to users. Secondly, the computer can be used as an access tool to obtain information from externally operated information storage and retrieval systems.

### **6.2. Development of automated Information Retrieval System**

Computers began to be used to assist the retrieval of the published literature and technical literature in early 1960s with the production of printed indexes. The purpose was to have faster access to current scientific and technical information. Subsequently, they were also used within the organizations for producing indexes and searching internal reports and other local collections of material. The radical change came in 1960s when System Development

Corporation and Lockheed Corporation began to purchase rights to magnetic tapes from producers and offer these tapes and search program through a telecommunication system that involved an inexpensive terminal and telephone connection. Any reference librarian with an access to a dumb terminal and a modem could call a local telephone number, use a simplified search language, and retrieve citations and abstracts from a number of indexes. H.P.Luhn of IBM introduced the concept of Selective Dissemination of Information (SDI) in 1961. In 1962, the producers of Chemical Abstracts decided to provide details of titles of documents available on magnetic tapes in order to enable the individual libraries to run their own SDI services. By early 1970s, when several of the published indexing and abstracting journals such as Biological Abstracts, Chemical Abstracts, Engineering Index, were produced by computer, several organizations acquired the corresponding magnetic tapes and used these to provide local information retrieval services, such as selective dissemination of information and current awareness bulletins. Later, the online public access catalogues developed by libraries became an efficient reference key to the library's collection. Because of indexing of the library's database it was possible to search this catalogue in many more powerful ways than manual search through card catalogue. These developments have paved for libraries to adopt computerized information retrieval practices.

However a major change in information retrieval systems occurred in late 1980s with the advent of the network facilities and access to the Internet and World Wide Web in early 1990s. The I.R. has changed considerably in the last decade with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR practices have become quite out-of-date which has led to the introduction of new IR techniques. Nevertheless, the basic concept and functions of I.R. are the same, though the procedures followed are different. Traditional finding aids to assist in information retrieval have focused on Subject Headings, descriptors, or established vocabulary. (Examples include LCSH, ERIC thesaurus, UMI's Controlled Vocabulary List). Items cataloged using

MARC format is one primary means to retrieve the item; In fact, the MARC format is an information retrieval standard: Z39.2. The subject headings are used to determine the primary focus of a book/article, i.e. to help precision if the user makes use of the established subject heading. The automated catalogs were created to use the traditional access points - author, title, series, subject and added entries. Over the years, adding more search capabilities to additional fields the online catalog capabilities have been enhanced. However in the modern information retrieval from the www digital pages, new concepts have been emerged such as query language, text and multimedia I.R., indexing and searching, parallel and distributed I.R., user interfaces and visualization, search engines, gateways and directories etc., and novel systems are being devised with the application of expert systems and artificial intelligence.

Thus the developments of automated information retrieval systems have been metamorphosed during the past fifty years and now these systems are more user-oriented than library-oriented.

### **6.3. Purpose of Information Retrieval**

In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant documents and discarding all the others. In a sense, this constitutes 'perfect' retrieval. Hence, the purpose of an automatic retrieval strategy is to retrieve all the *relevant* documents at the same time retrieving as few of the *non-relevant* as possible. When the characterization of a document is worked out, it should be such that when the document it represents is relevant to a query, it will enable the document to be retrieved in response to that query. When the retrieval system is on-line, it is possible for the user to change his request during one search session in the light of sample retrieval, thereby, it is hoped, improving the subsequent retrieval run. Such a procedure

is commonly referred to as *feedback*. The information retrieved can be classified in one of two categories of measurable statistics: whether the information retrieved is considered relevant, or whether all the relevant material was retrieved. These two metrics of recall and precision serve to express information retrieval performance.

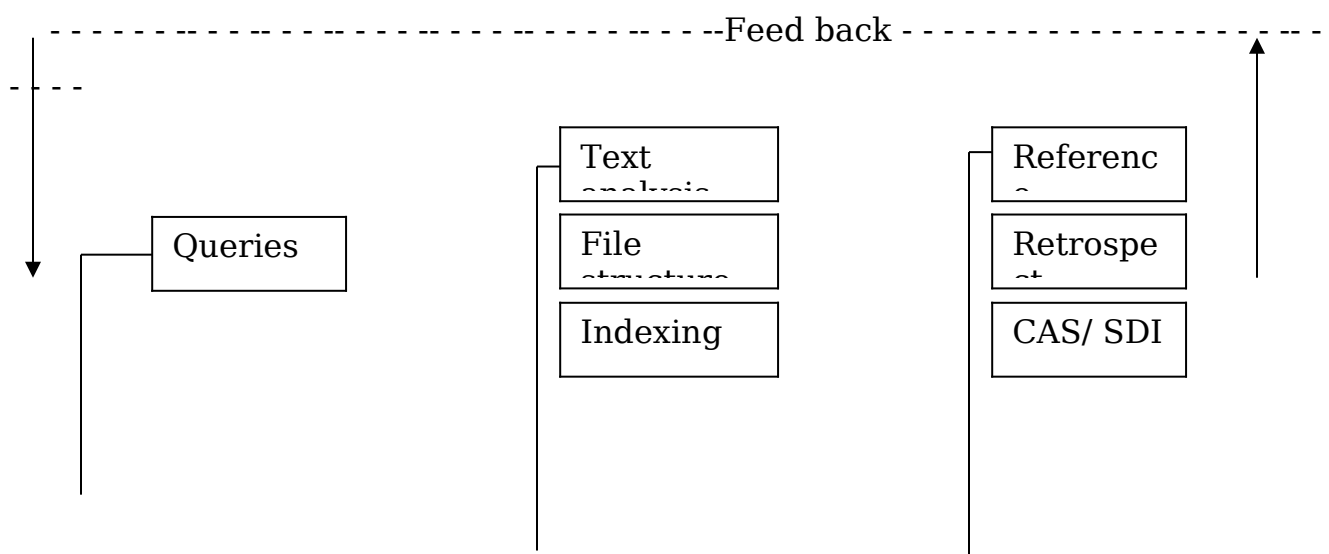
#### 6.4. Advantages of automated information retrieval system

Although the manual information retrieval system supports the information dissemination activities of libraries, its performance is limited with regard to accuracy and easy access. On the other hand, the automated retrieval system is more beneficial in terms of

- Improved customer service
- Improved productivity
- Decrease in enquiry time
- Overall improvement in library staff performance.

#### 6.5. Functions of automated information retrieval system

A typical IR system possesses the basic components such as: input (documents and queries), procedures (text analysis, file structures, indexing, searching), output (services) and feed back. The following bloc diagram illustrates the structure of a basic computerized information retrieval system.



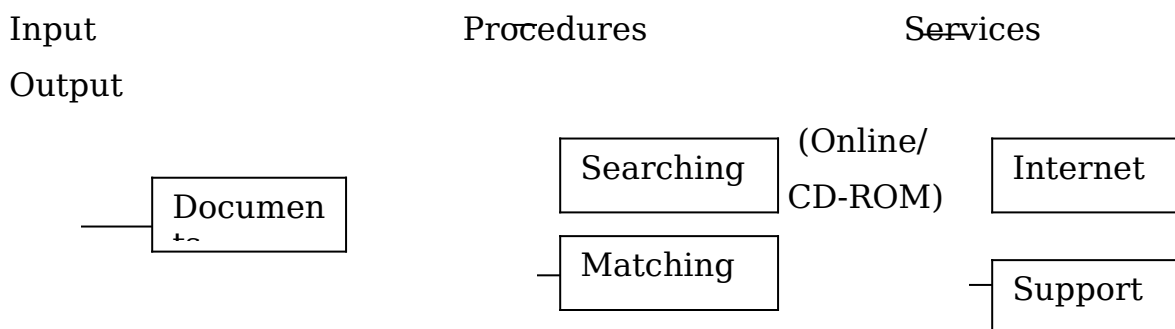


Figure 1. Structure of a Typical I.R. System.

**6.5.1. Input:** The two major components of information retrieval are users with queries and documents with answers.

**6.5.1.1. Users:** Users of automated IR system are expected to possess the computer access skills and search skills. It is difficult for a novice user to have the required skills hence libraries have to offer training programmes for users. The typical user in a computerized IR system requires (i) to have basic computer skills; (ii) understanding of how information is stored, indexed and retrieved in online systems; (iii) skills to access remote databases using search strategies. The online vendors have developed a number of 'front end' application software and on the Internet, the search engines and gateways are designed as more user friendly.

**6.5.1.2. Documents:** The present day library and information systems are hybrid in nature possessing print, and electronic resources and distributing the information following manual and computerized procedures. The information may be stored in printed books, serials etc, or may be d-stored in digitized form as magnetic, optical or online databases. The databases are two types, viz. reference or secondary databases like the card catalogues that retrieves bibliographic information; and full text that provides access to either abstract or

full text. The library staff as well as the users must be well aware of the document structure to access them with least effort.

**6.5.2. I. R. Procedures:** The common procedures an information retrieval system encompasses are as follows:

**6.5.2.1. Automatic text analysis:** Before a computerized information retrieval system can actually operate to retrieve some information, that information must have already been stored inside the computer. Originally it will usually have been in the form of documents. The computer, however, is not likely to have stored the complete text of each document in the natural language in which it was written. It will have, instead, a document representative, which may have been produced from the documents either manually or automatically. The starting point of the text analysis process may be the complete document text, an abstract, the title only, or perhaps a list of words only. From it the process must produce a document representative in a form, which the computer can handle. Luhn has contributed much to automatic text analysis. He proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements. The methods adopted for automatic text analysis are indexing, automatic key word classification, normalization of the text, bibliographic remarks etc.

**6.5.2.2. File Structures:** Any database will have several characteristics such as author, title, key word, ISBN etc. Depending on the type of document, the fields and descriptors may vary. For example a book may be identified by ISBN, while a serial by ISSN and a CD-ROM by its number. However, there are some fields like author, title, keyword etc that are common to any type of document. There are different ways of using file structures in document retrieval. Most of the file structures address themselves to applications in data management,

which is reflected in the terminology used to describe a document. For example the file structure using MARC, CCF or any other format is standard and logical in organization of document representatives or descriptors. This helps in data management and document retrieval. The file organization may be in the form of sequential files, inverted files, inverted sequential, tree, clustered files etc. that is suitable to an online environment.

**6.5.2.3. Indexing languages:** The language of the system is generally referred to as indexing language that could be based on controlled vocabulary or natural language (un-controlled). In controlled vocabulary the conceptual meaning of word(s) is encapsulated in a unique term with guidance to related terms like broad, narrow and related terms. However, the advent of the Internet open the doors for uncontrolled vocabulary and natural language indexing either from title or the complete text and these are more conducive to end user searching. The limitation here is they lack precision in retrieving relevant documents. Indeed the indexing language is the key component in matching the user request with the records of database through a search formula.

**6.5.2.4. Matching the request with Search Strategies:** The actual process by which the relevant information is located and retrieved is by searching the database, (reference or full text). When the search statement is executed the retrieval system retrieves the records, which are matching the search statement. The matching records are called 'hits' and the others are 'misses'. Searching online databases requires an information retrieval strategy. The search strategy is based on different kinds of planning: at the highest level a plan orders a linear and hierarchical planning for the request interrogation and a dynamic planning for the request modification. Traditionally, searching by simple terms, Boolean logic search, truncation weighted term are used as search strategies. The recent development of the system has allowed supplying some new judgments on this approach. The serial search, string searching, field

limit searching, cluster representatives, interactive search formulation and techniques of relevance feed back.

**Procedures:** The libraries may follow online or CD-ROM procedure to retrieve information from digitized information sources. Though the manual procedures are not completely replaced, majority of special and some university libraries are opting for CD-ROM and / or online retrieval.

**Online information retrieval:** There are thousands of databases available in almost all subjects of universe of knowledge that are developed by learned societies, educational and research institutions and individuals. There are online vendors like DIALOG, BLAISE et al who provides access to majority of important online databases. To make the search process easier, they have developed front-end software, which makes the access possible with least effort from users. However the access is subscription based. Now, the Internet is the largest store of databases that possess information on any concept or unit of thought. Online access to the Internet resources is possible interconnecting the computers through telecommunication links, either dial up or dedicated. The establishment of servers by Internet Service Providers makes the access more feasible. Though the users can have a direct access to online or CD-ROM information, the librarians still have a major role to play as a mediator. He has to guide the user in understanding logon procedures; selection of appropriate database out of the mass available; understanding Boolean and keyword searching; comprehending the search command language of software; knowing special indexes for databases etc. as time is money in online procedures.

**CD-ROM based retrieval procedure:** The use of online systems by library users is gradually being supplemented by public use of the systems as the library's users becomes more sophisticated and wants to do the searching by themselves. Vendors who previously offered only online index, abstract and full text services now offer their products in CD-ROM form. The libraries and information Centres with necessary workstations can operate these systems.

Publishers of reference works have also entered the process and many of encyclopedias, dictionaries and almanacs etc. are now available as CD-ROM products since one disc equals to many books.

**6.5.3. Services of computerized information Retrieval System:** The computerization of information retrieval system has brought in significant changes in the efficiency of traditional services and emergence of a variety of new services. The Internet has added advantage to the situation and the users of present day have more information that can be retrieved in real time with more accuracy, trespassing geographical boundaries. The modern information retrieval system offers the following services:

**6.5.3.1. Reference Service:** The computers were used in 1960s to retrieve information related to reference queries from abstracts and indexes. OPACs also help to answer queries related to literary output of a subject. However the Internet offers online reference service. The Global Reference Network of Library of Congress, available at [www.QuestionPoint.org](http://www.QuestionPoint.org) provides libraries with access to a growing collaborative network for reference librarians. Answerbank, Ask a Librarian etc. are some other online reference services aims to answer a wide range of queries on any topic.

**6.5.3.2. Retrospective services:** Retrospective search provides background information on a topic. Tedd described retrospective search as reverse of SDI where in the user query is run against document file as the user initiates search in retrospective search system. According to P.S.G.Kumar, the retrospective search system will perform the following functions:

- i. Translating the user query into a search strategy;
- ii. Comparing the search strategy with the descriptors of documents in the data base;
- iii. Posting of matched references to the users.

Computerized retrospective search systems were developed in U.S. in late 1950s, e.g. MEDLARS (Medical Literature Analysis and Retrieval System). Initially, the search was done in batch mode i.e., the queries of users were batched periodically and run against a related database. By the mid 1960s, online systems started and used for retrospective searches. In an online system the user can directly interact with the database, search and get the matches displayed on a VDU or can get a hard copy on printer. The online system is more advantageous than batch mode as it provides immediate access to database(s), so that reader can alter or improve his search strategy and thus increase the precision ratio.

**6.5.3.3. Current Awareness Services and Selective Dissemination of Information (CAS/SDI):** CAS is a system for reviewing the publication immediately upon receipt, selecting the relevant items to the research or works of an organization, and notifying the items that are matched to the users related with that task / information. In other words, it is an announcement mechanism to keep the users abreast of current developments in their field of interest. The aim is to bridge the time lag between the publication of a primary document and its coverage in an indexing/abstracting periodical. Different methods are adopted to retrieve current information services and SDI is one among them. You will learn about them in detail while studying more about information services. These services may be offered at local level as well as at global level. E.g. Local current awareness bulletins, Current Contents of Institute of Scientific Information, Chemical Titles of American Chemical Society etc.

#### **6.5.3.4. Internet - based information retrieval**

Today's Internet is a virtual cornucopia of information on just about any subject that could be imagined. But while the "Information Superhighway" has undoubtedly changed our lives forever, locating the exact material is difficult as the user weeds through mounds of related web sites. Hence the librarians have a

role to play by sifting the relevant information from irrelevant, repackage and serve to users. The Internet access offers a number of services such as

- Communication. E.g. E-mail.
- File or Document Transfer i.e. moving electronic documents, images, sounds etc. through file transfer protocol.
- Interactive browsing accessing databases of another remote computer system. E.g. Telnet is the tool to browse or search electronic resources.
- Bulletin Boards. E.g. USENET is a global bulletin board service.

In addition, the World Wide Web, a hypertext interface to information on web, opened the windows to world knowledge and 'global village' became a reality with access to information from any where in the world. The Internet facilitates use of a remote online catalogues and databases, identify material on a topic of interest using different search features.

#### **6.5.3.5. Support services:**

The automation of information retrieval facilitates support services such as inter library loan and document delivery, as the copies of the document can be sent electronically or as FAX. Further, automatic machine translation services are also available.

**6.6. Feed back:** The word feedback is normally used to describe the mechanism by which a system can improve its performance on a task by taking account of past performance. In other words a simple input-output system feeds back the information from the output so that this may be used to improve the performance on the next input. The feed back from users helps in evaluating the profile performance and identification of modifications needed in the profile and updating the profile. The notion of feedback is well established in automated information retrieval systems and used with considerable effect.

**6.7. Evaluation of information retrieval system:** Evaluation of information retrieval system is necessary to ascertain the performance of system and to make necessary modifications. Many efforts have been made to develop solutions to the problem of evaluation of I. R. systems and to measure the effectiveness of the system. For e.g. if we would like to assess, what benefit will

users obtain (or what harm will be done) by replacing the traditional sources of information by a fully automatic and interactive retrieval system? It is hard to interpret without proper study. Clevedon listed six main measurable quantities:

- The *coverage* of the collection, that is, the extent to which the system includes relevant matter;
- The *time lag*, that is, the average interval between the time the search request is made and the time an answer is given;
- The form of *presentation* of the output;
- The *effort* involved on the part of the user in obtaining answers to his search requests;
- The *recall* of the system, that is, the proportion of relevant material actually retrieved in answer to a search request;
- The *precision* of the system, that is, the proportion of retrieved material that is actually relevant.

**6.8. Summary:** the early efforts of automation are stand-alone functions – like online searching, interlibrary requests etc. As libraries approach the next generation of automated library functioning in 1970s, integrated approach to information retrieval using OPACs and other internal databases increased. As discussed in this unit, though the initial application is restricted to internal information retrieval, the 1980s had witnessed the application of I.R. techniques for accessing external databases like MEDLARS, AGRINDEX, and CHEMICAL ABSTRACTS etc. However the emergence of the Internet in 1990s metamorphosed the situation and paved new ways of information retrieval, an instantaneous, real time, global access to information. Although the present day information retrieval systems are end-user oriented the librarians still have an important role to play in selecting relevant databases and serving the user with information that has high relevance and precision values.

## **6.9 Further Readings:**

1. Rowley, J.E. Computers For Libraries, 2<sup>nd</sup>ed. London, Clive Bingley, 1986.
2. Lovecy, Ian. Automating Library Procedures: a survivor's handbook. New Delhi, D.K.Agency, 1984.
3. Wright, Keith C. Computer-related technologies in library operation. Hampshire, England, Gower, 1995.
4. Tedd, L.A. An introduction to computer based library systems. 2<sup>nd</sup> Ed. New York, USA, John Wiley & sons, 1984.
5. Kumar, P.S.G. Information Technology Applications (Theory & Practice) New Delhi, B.R.Publications, 2004.

**6.10. Unit-end-questions:** What is the purpose of an automated Information Retrieval system?

1. What are the functions of an Information retrieval system?

## **6.11. Answers to unit-end questions:**

1. The primary purpose of an information retrieval system is to retrieve all relevant documents on a subject, restricting the retrieval of irrelevant documents following an automated retrieval strategy.
2. The basic functions of an I.R. System includes input (documents and queries), procedures (text analysis, file structures, indexing, searching), output (services) and feed back. The application of computers makes the functioning of the system more effective with high retrieval rate, more accuracy, and less time and manpower.

