UNIT 3 INFORMATION RETRIEVAL SYSTEMS

Structure

3.0 OBJECTIVES

After reading this unit, you will be able to :

- understand the definition of information retrieval systems;
- know the theoretical foundation and models of information retrieval systems;
- get yourself acquainted with design and operation of IRS; and
- explain the method of searching information from IRS.

3.1 INTRODUCTION

It was Calvin Mooers who in 1950 coined the term “information retrieval” and described it as “searching and retrieval of information from storage according to specific subject.” The word retrieval means to discover and bring to the notice of the users the documents in which information is embedded. Again B.C. Vickery has described it as “ retrieval is essentially concerned with the structure of the operation of the device to select documentary information from the store of information in response to several questions”

The retrieval systems are usually in a state of continuous gradual revision; data are added or withdrawn; new index points inserted; syndetic relationship changed. The development of effective retrieval technique has been the core of IR research for more than 30 years. Nowadays multimedia indexing and retrieval techniques are being developed to access image, video and sound database without text descriptions.

The information retrieval system is certainly not a new concept; it is an integral part of the communication process, a direct outgrowth of the desire among men to communicate with eachother.

- The classification of retrieval techniques that has been proposed by Hicholas Belkin and Bruce Croft are:

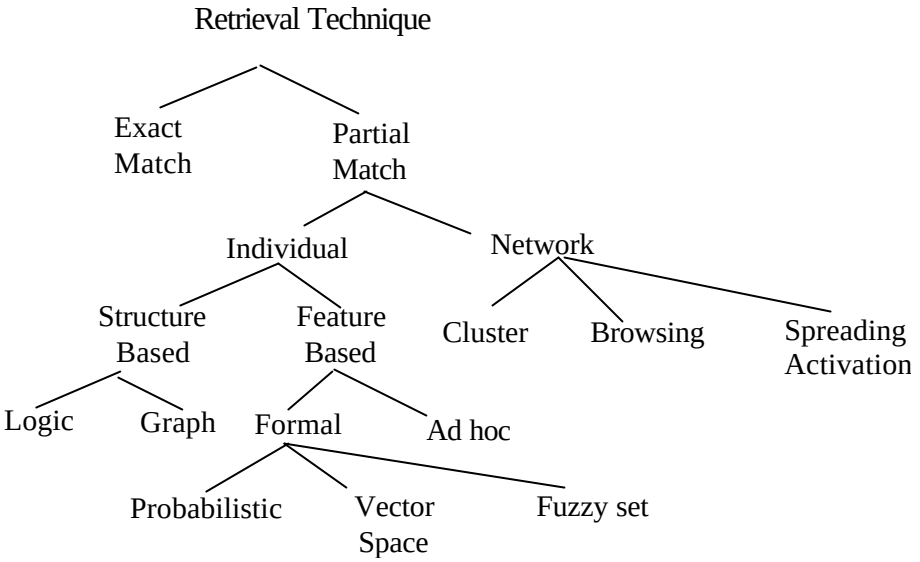


Fig. : Classification of Retrieval Techniques

Belkin and Croft distinguish between exact and partial match techniques. Exact match techniques are currently in use in most of the conventional IR systems. Queries are usually formulated using Boolean expression and the search patterns within the query have to match with exactly the text representation of the document to be retrieved. Partial match retrieval technique as opposed to exact match technique is categorised into individual and network. Individual techniques search single document nodes without considering the document collection as a whole. In the feature-based techniques, documents are represented by sets of features or index terms. The index can be either defined manually or be computed automatically. In structure-based techniques, documents are represented in a more complicated structure than just a set of index terms as used for the feature based techniques.

In network based methods, the set of all documents and their relationship are used to find the most relevant documents. With this method, the technique query. In clustering, most similar documents are clustered together and all documents are grouped into a cluster hierarchy until a ranked list of lowest level clusters are produced. Spreading activation is similar to browsing. From the start nodes, other nodes connected to that node are activated. Activated nodes then propagate or spread themselves through the network.

Theoretically there is no constraint on the type and structure of the information items to be stored and retrieved with the information retrieval (IR) system. Until recently information retrieval systems were limited to searching textural information. Gerard Salton has defined an information retrieval system as a “system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations.”

According to Alken Kent , any information retrieval system entails a series of processes or steps, which are as follows:

i) Analysis involving perusal of the record and the selection of point of view (or analytics).

ii) Terminology and subject heading control involving establishment of some arbitrary relationships among, 'analytic' in the system.

iii) Recording the results of analysis on a searchable medium.

iv) Storage of records or source documents, involving the physical placement of the record in some location.

v) Question analysis and development of search strategy involving the expression of a question or a problem.

vi) Conducting of search involving the manipulation or operation of the search mechanism in order to identify records from the file.

vii) Delivery of results of search involving physical removal or copying of a record from files.

Thus, any information retrieval system has three components - input, process and output. The storing of information is the input component. Generally the search or retrieval of information from the information retrieval system is through a query processing system. The information stored in the system is indexed using some indexing technique using key words. The processing system matches the key words of the query language with that of the key words under which the information items have been indexed. The matching results into the response output which may be the answer to the user in response to his request or search for information.

3.2 THEORETICAL FOUNDATIONS

The development of various techniques to retrieve information has been a major area of research interest and has been renewed from time to time through greater emphasis on computerised information retrieval systems. The examples of early theoretical approaches to are classification theory; linguistic theories in the context of automatic indexing; psychological approaches and the early structural models of Fairthrone and others. Any information retrieval system is based on some theory. Theory is a set of sentences in a formal language with a few powerful axioms, some special rules of inference and a rich body of true theorems that captures the essential phenomena and concepts. Taking "theory" in its widest sense, any one setting up a retrieval system must have some theory relating to the function of the system. In absence of any general accepted theory, any formulation that appears to deal with or relate to any part of the storage and retrieval process is potentially a part of the theory of information retrieval.

Swets regarded the retrieval process as having two stages. In response to a request, the system first calculates for each items of information the value of search functions. This function discriminates between relevant and non-relevant information because its distribution for relevant information is different from that for non- relevant one. The system then selects those items whose match values are highest or higher than a certain threshold. The classification of retrieval technique as part of theory is already discussed in the introduction of this Unit.

As early as 1963 Swets developed an evaluation model based on statistical decision theory. The first book on the theory concerning information appeared in 1961

describing the principles of index construction or subject description of documents. The most important application of a concept from logic was the application of Boolean lattices to logical combinations of descriptors. Another important development was Shannon's information theory to indicate desirable statistical characteristics of index terms. There have been theoretical approaches to IR from the viewpoint of the function or functions which the system performs. The performance of a system must be explicitly stated in any theory. While any retrieval system must be based on some theory of retrieval, such implicit theories are extremely difficult to extract or analyse. Even some explicitly formulated theories are formulated in such general terms, with such loose connection between the theory and system design, that they are difficult to evaluate. There are theories relating to the relevance feedback and manipulation of wide terms. The "Weighing function" formation 'of Robertson and Sparck Jones can be mentioned. Then concerning the indexing and retrieval effectiveness the important theoretical contributions are by Marton, Kuhns and Cooper. the probabilistic and utility theoretic indexing by them is worth while to mention. Saltons' theory of indexing is another important theoretical development in the field of information retrieval. Attempts are on in the direction of building an integrated general theory of information retrieval.

3.3 MODELS OF INFORMATION RETRIEVAL SYSTEMS

1.3.1 Models Based and Input and Output

Different models of information retrieval system can be recognized, based on input and output aspects. We can group them into 4 basic models viz.

1) Data Retrieval Model

2) Information Retrieval Model

3) Knowledge Retrieval Model

Data Retrieval Model

A data retrieval model calls for the organizational structure of the content (data) based on various criteria such as, properties of population, clusters and other entities. A data retrieval model essentially handles data which can be taken as unprocessed information. There are a number of economics related data retrieval systems providing various types of socio-economic data. The census is a data retrieval system. Similarly, data available from national survey organisations and central statistical organization can be taken to be a numerical data system. The information retrieval systems based on data also retrieve information. The expression of information, thus, needs be very precise. In this context the data retrieval model does a simple model of information retrieval need specific matching technique, viz., a taxonomic structure of various entities involved and their properties.

Information Retrieval Model

Information is data processed and oriented to a purpose. It actually combines several data into a relational structure. Information retrieval is, therefore, a more complex model. It has to comprehend generally multidimensional relationship. It is not amenable easily to a taxonomic structure. The representation of information may be based on

a relational database structure using some associative mathematics. The expression of information need is complex and time consuming. It draws out for a long conversational or browsing process and the informational retrieval model must incorporate such facilities of interfaces.

A distinction is made between data retrieval and information retrieval. In practice the distinction is one of degree rather than kind . Figure 2, shows a spectrum of various attributes of data retrieval and information retrieval:

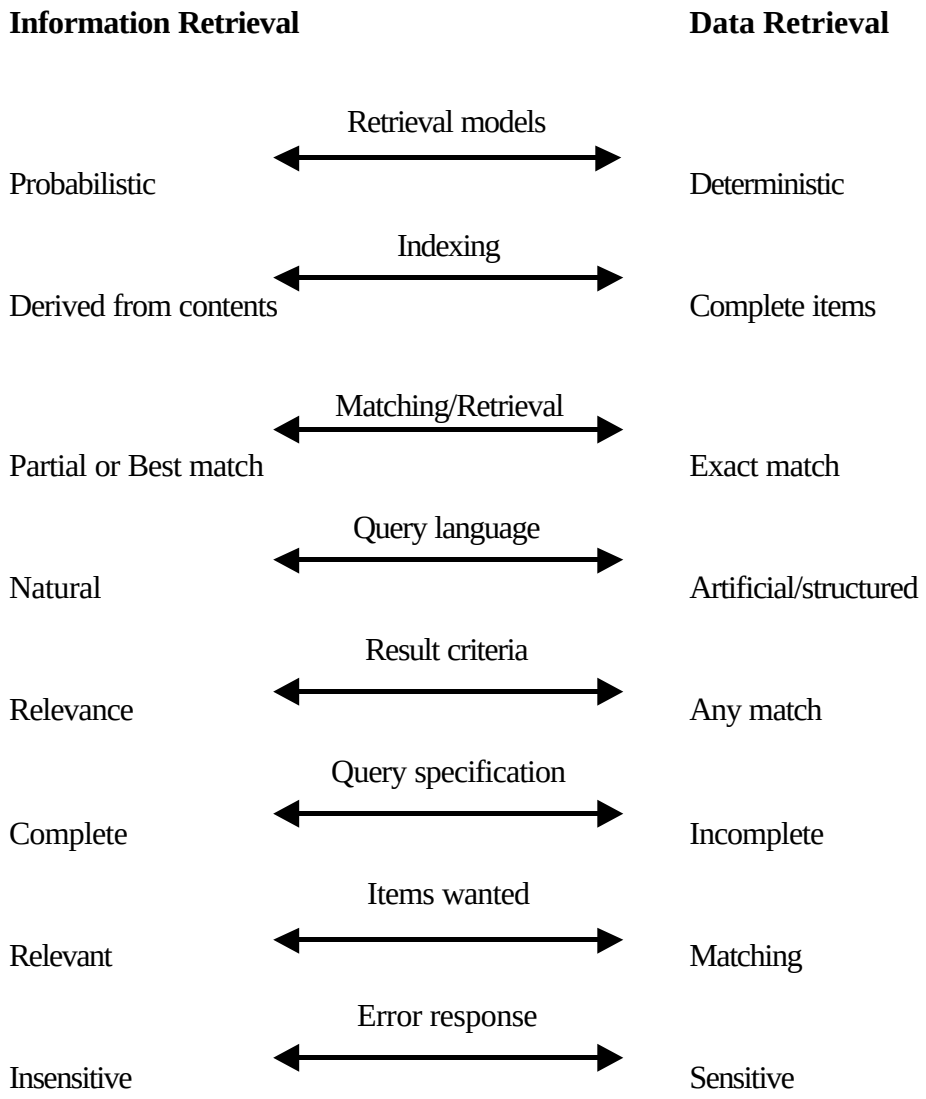


Fig. 2: Information Retrieval/Data Retrieval Spectrum

In information retrieval, the underlying model of providing access to information is probabilistic. It is concerned with such subjective and indeterminate issues as whether (and to what degree) a document satisfies a user's need for information i.e. whether it is relevant for that user's request. Data retrieval on the other hand is deterministic with regard to retrieval operations. Most information retrieval systems derive their index element from the content of the items. The derivation may be simple extraction (such as extracting key words from a text), inferential extraction or it may be intellectual analysis and assignment of index items. In data retrieval, the element itself, in its entirety, is the indexing unit. Obviously this spectrum is not really smooth or continuous since both types of indexing may be present within the same system.

In the actual retrieval, the algorithms used for matching the query and the index elements are based on the particular retrieval model. Information retrieval model leads us to a class of retrieval algorithms that are probabilistic in nature, and may involve the actual calculation of probabilities and use of statistical inference methods, or may take another approach based on another model of the document space (such as Salton's vector space model), They attempt to find all of the potential (partial) matches between query and document and to rank them based on some measure of "goodness" so that the best matches receive the highest rank . The data retrieval algorithms are deterministic, and therefore demand an exact match between the query specification and the contents of the database.

The queries in information retrieval systems are commonly expressed as a natural language statement of the searcher's needs for information. These queries are inherently imprecise and may be ambiguous in many cases. In data retrieval, the query is usually expressed in some sort of structured query language with precise syntactic and semantic characteristics. The query types, thus, reflect the underlying model of the retrieval system.

Knowledge Retrieval Model

Knowledge retrieval models are based on the computational reasoning technique developed in Artificial Intelligence (AI) research. They are based on the knowledge of human experts. In order to facilitate decision-making and problem- solving, intelligent knowledge-based information retrieval models are being designed. The three major components are :

i) Knowledge base

ii) Inference Engine

iii) User interface

The knowledge retrieval model is the most sophisticated model of information retrieval,

3.3.2 Models Based on Theories and Tools

Based on theories, tools and techniques a number of other models of information retrieval systems have been developed on contemplated. These are:

i) Linguistic Model

ii) Mathematical Model

iii) Psychological Model

iv) Economic Model

Linguistic Models

In linguistic model of information retrieval, we can study the system from the point of view of the properties of the language. The various ways of storage of information are eventually based on natural language. In short, the language used has three functions:

a) It represents the content of documents and other forms of information.

b) The information problem in represented in terms of language.

c) Language is used in computer processing and also in searching and retrieval of information.

The language works on three base :

i) Semantic base which conveys meaning from one human being to another.

ii) Syntactic base which helps in the formation of semantics by the use of grammar, and

iii) The vocabulary which supplies different meanings to terms for the formation of explosion, expressed in sentences, paragraphs etc.

The logical structure of a language and the taxonomy of the language refers to the relationship between vocabulary and concepts. The vocabulary generally refers to the logical structure. The vocabulary control indents thesaurus content and technical glossary control. The indexing language with control of expression terms provides the basic model for information retrieval. Use of associative mathematics in search logic and in search expression formulation provides yet another type of language control in information retrieval.

Mathematical Models

Mathematical models are essentially based on representative mathematics as well as associative connections. In particular, cluster analysis and clustering techniques are used on an experimental basis in automatic abstracting and indexing. Use of set theory and Boolean logic is a familiar method of mathematical modeling of information. Concept of similarity measures and choice of variable and the combinational aspect of clustering tries to provide semantic structure for information represented.

Psychological Model

The psycholinguistics approach to information retrieval led to the study of the formation of concepts in human mind, the way in which the human thinking process arranges the ideas and present it at the time of inquiry and the types of retrieval it demands while searching. The studies of Belkins, Brooks and Oddy on anomalous state of knowledge provides an interesting insight in relation to the information retrieval process. Further, studies in information retrieval and artificial intelligence have thrown significant input bringing in a harmonious coupling of psychological theory with information retrieval.

The Economic Model

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

User Model in Information Retrieval

In IRS, the user model is used to provide assistance to the user in the query formulation process. The goal is to express the information requirements in the best possible way. Clearly the best way is the one that provides the system with enough

input information to retrieve all relevant documents. Frequently users have a hard time specifying explicitly what exactly they are looking for. It is the task of the user model component of the IR system to automatically help and complement the user's interests based on their previous search behaviour.

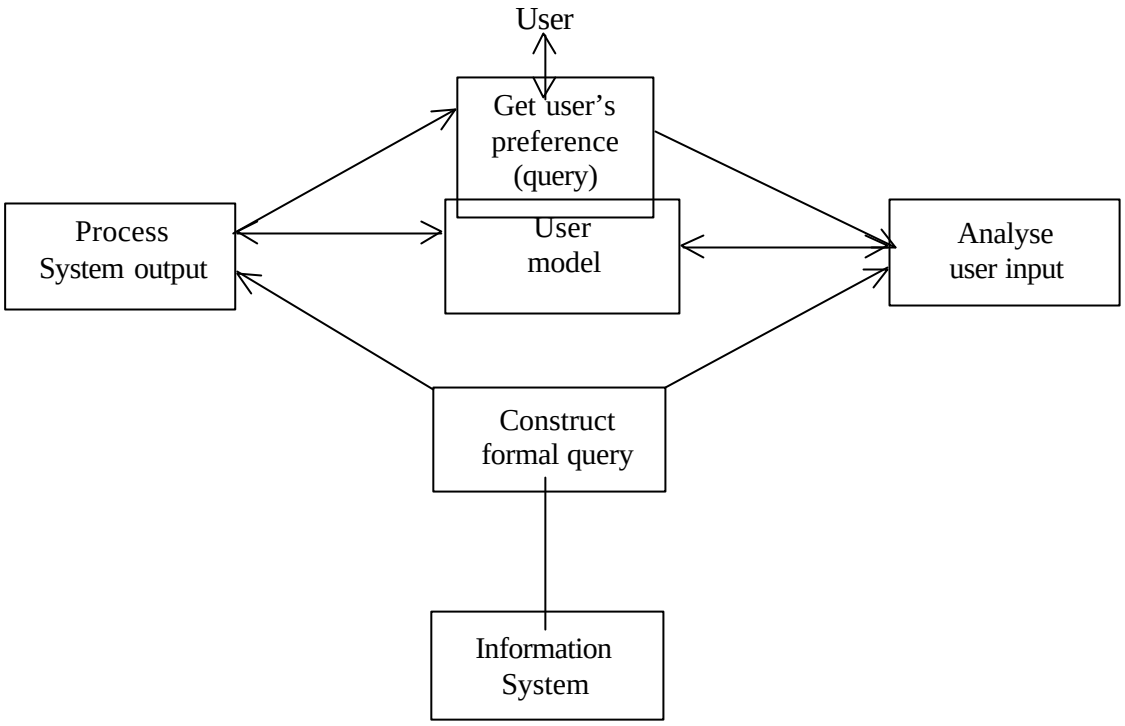


Fig. 3: User Model in Information Retrieval System

Figure 3 illustrate the central position that a user model can assume in information retrieval system. An IR system enhanced with user modeling techniques will normally start by getting the user's preferences. For example, there can be a statement of the user's interests as a self-description, or it may be a SQL based query. This input is subsequently analysed, using the user model, and the user model is updated accordingly. Then the formal query is constructed and processed, based on the user's preferences. Afterwards in close interaction with the user mode, the output is prepared for presentation to the user and the user model is refreshed. Finally, the user can evaluate the query and restart the whole cycle again if needed.

3.4 IRS DESIGN AND OPERATION

The overriding principle underlying any IR system is that it be governed by the laws of the host organization it serves. Some of the important factors to be considered in the design of IRS are :

i) Specify the user class to be served. Distinguish between direct users and beneficiaries. If there is more than one user or beneficiary class for the same IR systems, assign and announce clear cut priorities on the basis of matching the characteristics of the system with those of the user class.

ii) Specify the uses and problem-class for which the system is intended, and fix priorities among them, although designing an IR system with a defined problem is a difficult task.

iii) Specify the range of information items which should be acquired, organized and stored in the system's database.

iv) Cross-validating and checking inputs to the database.

v) The system should be cost effective to test, restructure, and to edit the database for significance, relevance, validity, and redundancy (may be periodically or when flaw occurs).

vi) The hardware software and other components. It is essential to analyse the hardware and software requirement of the system.

Design

The design of an IR system involves problems of analysis and problems of synthesis. The objective of analysing an IR system is to arrange a set of feasible IR systems in a partial ordering. The IR system should be effective and efficient in its functioning. An IR system, to be viable, should conform with the basic principles governing the processing of information in the host institution.

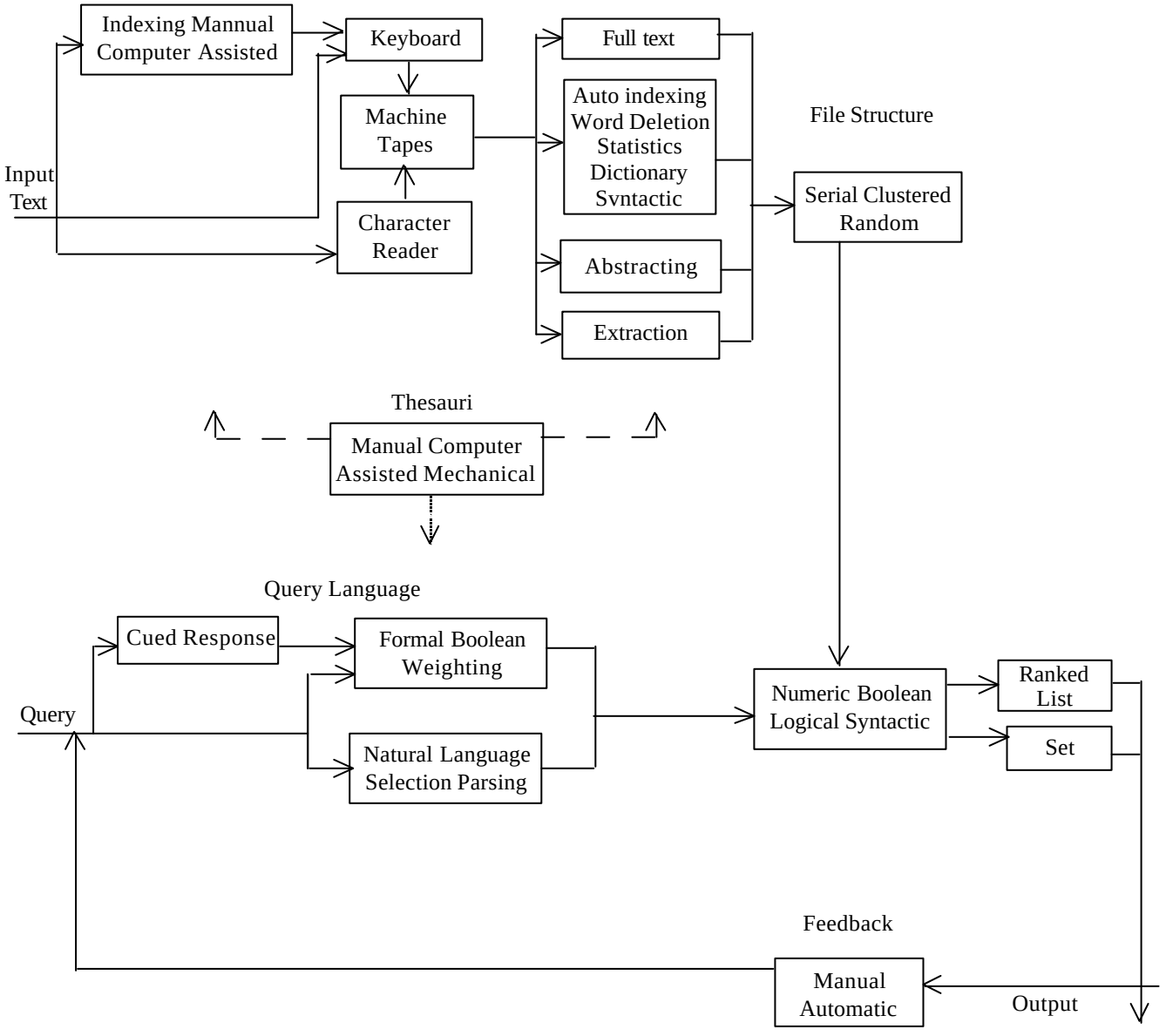


Fig.: Information Retrieval System Components

3.5 SEARCH STRATEGY

Basic Search Techniques

In a bibliographical information retrieval environment, searches can be divided into two main classes-known item search and unknown item search. A known item search is what is conducted when the user knows something about the item being sought. This may be any key, such as author, title, publisher, ISBN, and so on. An unknown item search is conducted when 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 such an item exists that can meet their information requirements. There are different types of searches which are helpful to understand the entire process of search strategies.

Keyword and Phrase Search

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

In a phrase search, the system searches for the entire phrase rather than each individual key word forming the phrase. Phrase searches can be conducted only in those fields that are phrase indexed. If the index file comprises only single terms, then phrase search cannot be conducted, unless proximity operators are used whereby the system will searche for each constituent keyword in the search expression separately, and retrieve only those records where the keywords occur consecutively. A search phrase can simply be entered through the keyboard, or selected from an index file or vocabulary control tools like subject headings lists and thesauri.

Different search systems provide different facilities for conducting key word and phrase searches. For example, in a Dialog search one can simply enter a key word or a phrase preceded by the search command. The user can restrict the search to one or more fields.

Many bibliographical information retrieval systems provide two types of search facilities for conducting an unknown item search; keyword search and subject search.

A keywords search allows users to enter one or more key words pertaining to their query. These keywords can be chosen by the user in any combination depending upon the requirements, and there are several search operators that can be used to combine several keywords to formulate a search expression. The search keywords can appear anywhere, or in one or more chosen fields, in the database records. A subject search allows the user to submit a subject expression that reflects his or her information requirement. Such a search is conducted on the subject field that contains the subject headings assigned by the indexes when the database was created. Thus, a record will be retrieved only when the user's subject search expression exactly matches the subject heading assigned by the indexes. For standardizing the process, and also helping the user identify the appropriate subject headings, IRS uses certain tools, called vocabulary control tools.

Boolean Search

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

The AND search allows the user to combine two or more search terms using the Boolean AND operator. The search will then retrieve all those items that contain all the constituent terms. For example, the search expression "Internet AND computer" will retrieve all those records where both the terms occur .The search is restricted by adding more search terms. The more search term are AND ed, the more restricted, or specific will be the search and as a result the smaller will be the search output. Sometimes, a search may produce a blank result if too many search term are AND ed.

Truncation

Truncation is a facility that enables a search to be conducted for all the different forms of a word having the same common root . As an example, the truncated word COMPUT* will retrieve items like COMPUTER, COMPUTING, COMPUTATION, COMPUTE, etc. A number of different options are available for truncation viz. right truncation(as in COMPUT* example), left truncation, and making of letters in the middle of the word. Left- Turn truncation retrieves all words having the same characters at the right - hand part e.g.*HYL will retrieve words like METHYL, ETHYL etc. Similarly middle truncation retrieves all words having the same characters at the left- and right hand parts. For example, a middle truncated search term COL * will retrieve both the terms COLOUR AND COLOR.

Proximity Search

This search facility allows the user to specify :

1) Whether two search terms should occur adjacent to each other,

2) Whether one or more words occur in between the search terms,

3) Whether the search term should occur in the same paragraph irrespective of the intervening words, and so on.

The operators used for proximity searching and their meanings differ from one search system to another. The various types of proximity search facilities and the corresponding operators are available in CD-ROM and online database.

Field-specific Search

A search can be conducted on all the fields in a database or it may be restricted to one or more chosen fields to produce more specific results. Specific fields and codes vary according to the search systems and database.

Limiting Search

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

Limit	Qualifier	Example
English - Language document only	/ENG	SELECT URBAN (s)CR IME?/ENG
Patents only	/PAT	S TRANSISTOR?/PAT

Range Search

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

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

Search Tools :

Library and information professionals have since been using four types of tools for organizing information. They are:

1) **Classification Schemes**

The classification schemes such as, Dewey Decimal Classification (DDC) Universal Decimal Classification (UDC) Library of Congress Classification (LC), Colon Classification and so on, are used for classifying documents, organization files and also for the physical organization of documents in libraries.

2) **Catalogue Codes**

The catalogue codes, such as Anglo-American Cataloguing Rules, Classified Catalogue Code, etc., are used to prepare catalogue records of documents, which provide information to a user about what a given library/information center possesses.

3) **Standard Bibliographic Record Formats**

Standard record formats such as ISBD and MARC (Machine Readable Cataloguing) formats are used to prepare machine readable records of bibliographic and other types of documents.

4) **Vocabulary Control Devices**

Vocabulary control devices such as thesauri and subject headings lists are used to standardize the terminology, which can be used both at the time of indexing and searching records.

All these tools can be used for organizing information in various types of information systems including digital library systems. However, these are only basic search tools which may be used and there are many more search techniques available for specific information retrieval systems.

3.6 EVALUATION OF IRS

Any information system exists to provide the seeker of information the document which bears the information or answers his query: The evaluation is a diagnostic activity to understand the performance of a system. It reveals the strength as also the weakness of an information system. It informs about the social benefits that accrue from the system. It also tells us about the economic aspects of the system, such as cost various aspects etc. On the basis of a careful evaluation one can thus ways for improving the system, if required. Evaluation is rightly called an investment for the future.

Evaluation Methodology

The evaluation programme of an information system involves a number of distinct steps. Let us understand these steps:

1) The first step is to be clear about the scope of evaluation. That is to say the purpose of evaluation should be very clearly defined. The scope should be defined precisely before the designing and execution of an evaluation programme.

2) The second step is the designing of the evaluation programme. The design should be such so that it suits the objectives and purpose defined earlier. The success of the evaluation programme depends upon the choice of appropriate design.

3) After deciding about the scope and design of the evaluation programme, the next step is the execution proper. The execution includes the collection of data, its organization, analysis and, lastly, the drawing of conclusions.

4) The fourth step is to analyse the conclusion and the interpretation of the results.

5) The fifth and the final step is to modify the information system on the basis of the result of evaluation as revealed in steps 3 and 4.

Irrespective of the methodology followed, the purpose of evaluation of any information system is to find out how well the input performs and what measures need be taken for its improvement. Sometimes, the evaluation of a particular IRS may provide a clue for the design and development of other systems.

Criteria for Evaluation

The criteria on the basis of which an IRS can be evaluated are:

1) Recall and precision and related factors affecting retrieval efficiency

2) Cost

3) Response time

Recall and Precision

The effectiveness of information retrieval can be measured by the ability of that system to retrieve the relevant documents and hold back the irrelevant ones in a given collection in relation to a particular query. The ability to inform about the retrieval of relevant documents and withhold the irrelevant ones are called recall and precision powers of the system respectively. Though theoretically 100% recall and precision is desired in practice it is not possible, as these two factors are inversely proportional to each other. The system in which these two factors are at the optimum level will be regarded as the best one and would be preferred for application.

In response to a query, all the relevant document may not be retrieved in a search, only a part of them may be retrieved. Similarly all the documents retrieved may not be relevant, though a number of non-relevant documents also remain as not retrieved. This can be illustrated in the following formats:

Document	**Retrieved**	**Non retrieved**	**Total**
Relevant	a	b	a+b
Non- relevant	c	d	c+d
Total	a+c	b+d	a+b+c+d

Recall is the retrieval of relevant documents by the system. Recall ratio can defined as the ratio of the number of relevant items retrieved to the total number of relevant documents in the system. This can be mathematically represented as :

$$\frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}} \times 100$$

$$\text{or. } \frac{a}{a+b}. \times 100$$

Suppose there are in all 100 relevant document in a file and the index is able to retrieve only 75 of them and misses 25, then the recall ratio is 75/75+25×100=75%

Precision Ratio

Precision is the capacity of the system to withhold non-relevant document. Precision ratio may be defined as the ratio of the relevant retrieved documents to the total number of documents retrieved from the file. Mathematically it may be represented as:

$$\frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}} \times 100$$

$$\text{or. } \frac{a}{a+b}. \times 100$$

Suppose the total number of documents retrieved are 150, out of these 75 are relevant, then the precision ratio is

$$\frac{75\times100}{150} \text{ or } 50\%$$

Many a time it is difficult to know the actual number of relevant documents in the store. Nevertheless, findings of recall and precision are helpful in assuring the quality of I.R.S.

Besides recall ratio and precision ratio, the other relevant measures which provide the retrieval efficiency of a system are:

1) Noise ratio.
2) Fallout ratio.
3) Novelty ratio.

Noise Ratio

It is complementary to the precision ratio. It shows the numbers of non- relevant documents out of the total documents retrieved. Mathematically it can be represented as:

$$\frac{\text{Total No. of non-relevant document retrieved}}{\text{Total No. of document retrieved}} \times 100$$

The lesser the noise ratio the more efficient a retrieval system will be.

Fall out Ratio

It shows how many non-relevant document, out of the total number of document in the store have been retrieved by the retrieval system. Mathematically it may be put as:

$$\frac{\text{Total No. of non-relevant document retrieved}}{\text{Total No. of document in store}} \times 100$$

Novelty Ratio

It is the proportion of nascent or new information items, which the system is able to bring to the attention of information seekers for the first time. Out of the total number of relevant document, a small percentage may be of such documents which contain nascent information. If out of the 100 relevant documents there as 15 such documents the Novelty Raito will be 15% i.e.

$$\text{Novelty Ratio} = \frac{15}{100} \times 100 = 15\%$$

An efficient retrieval system will bring to the attention of the user more of such documents which provide novel or new or nascent information.

Indexing Exhaustivity

The exhaustivity of a system refers to the accuracy and depth with which the various concepts contained in the system are covered. Exhaustivity is the property of index description. The indexing exhaustivity is connected with recall power of the system. A system having high indexing efficiency possess high recall power.

Cost

Cost is an important factor of IR system evaluation. Cost may relate to initial expenditure required to develop a system and also other direct charges, concerned with manpower, material, tools and other initial costs. The cost is a composite factor which also includes the effort involved on the part of the indexer and the time involved in the preparation of index and also the search time and search efforts on part of user. Initial cost can easily be measured but the cost of effort would be matter of experience and realization. If a particular system is less costly than the system it better than the other. The case of the use of the system by the user can be related to this aspect.

Response Time

Response time is another important factor for measuring the efficacy of the system. Response time should be measured while the users are interacting with the system.

If a system requires less time to retrieve information it would be economic and would be better than the other taking a longer time to retrieve the same information.

Self Check Exercise

1) Discuss the various information retrieval techniques.

2) Name the information retrieval models based on tools and techniques.

3) Explain search strategy in brief.

4) Discuss the criteria for evaluation of IRS.

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

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

.....

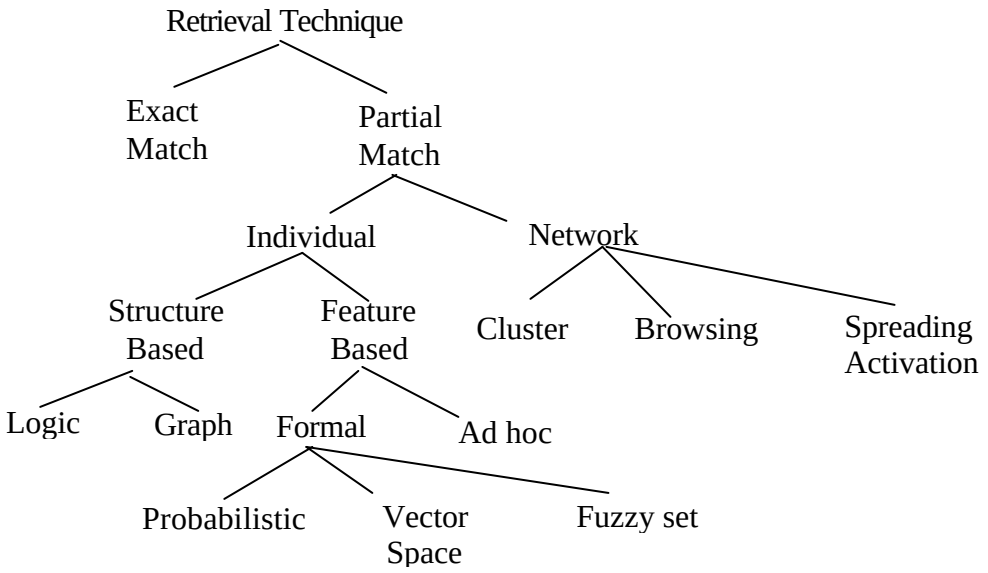
.....

3.7 SUMMARY

The development of effective retrieval techniques has been the core of IR research for more than 30 years. A number of measures of effectiveness have been proposed. Effective interfaces for text based information systems are a high priority for users of these systems. With the increase in the use of the internet, there has been a corresponding increase in the demand for information retrieval system that can work in wide area network environments. Search engines like INFOSEEK, LYCOS, etc., index web pages and provide access to them. Developing databases and providing search and retrieval access in an integrated manner have been the most important aspect of developing IRS. The technique for indexing and query optimisation have been the major issues.

3.8 ANSWERS TO SELF CHECK EXERCISES

1) The information retrieval techniques can be identified into exact match and partial method:



The exact match techniques are currently in use in most of the conventional IR systems. Partial match retrieval technique is opposite to exact match technique. While in the feature based techniques the information is represented by a set of features or index terms, in the structure based technique the documents are represented in the a more complicated structure than just a set of index terms.

2) The information retrieval models are :

   i) Linguistic Model

   ii) Mathematical Model

   iii) Psychological Model

   iv) Economic Model

3) There are 5 types of search strategies depending upon the kind of queries:

   i) Boolean search

   ii) Matching Function

   iii) Serial Search

   iv) Cluster Based Retrieval

   v) Interactive search formulation

Boolean search is a technique that combines search terms according to Boolean logic (operators). Three types of Boolean search are possible : AND search; OR search and NOT search. The search is basically through key words or phrases. The key word search is the simplest form of search facility offered by a search system. In key word search mode, the system searches the inverted file (index) for each key word forming a search expression. Search expressions may make use of either Boolean or proximity operators.

e.g. 'WATER' AND 'POLLUTION'

'POLLUTION' AND (WATER OR SEA OR RIVER)

'POLLUTION' AND (WATER OR SEA OR RIVER)

AND NOT SEWAGE

One can also go for proximity search and field specific search.

4) The criteria on the basis of which an IRS can be evaluated are :

   i) Recall and precision

   ii) Fallout

   iii) Cost novelty

   iv) Response time

3.9 KEYWORDS

Boolean Search : Developed by George Boole, It is a search strategy for formulating search query expressed in terms of index terms (or key words), combined by the use of logical operators AND, OR, and NOT.

Feedback : The mechanism by which a system can modify/ improve its performance of a task by taking account of past performance.

Information Retrieval Systems : The system to store items of information that need to be processed, searched, retrieved and disseminated to various user population.

3.10 REFERENCES AND FURTHER READING

Cleverden C.W. (1972). On the inverse relationship of recall and precision. *J of Documentation*. 28(3); 195-201.

Doyle, L.B. (1975). Information retrieval and processing. Los Angels : Melville.

Fairthorne, R.A. (1961) The mathematics of classification, towards information retrieval. London: Butterworths. 1-10.

Foskett, A.C. (1977). The subject approach to information, London : Clive Bingley.

Information storage and retrieval system (1997). IGNOU course material. MLIS-03. New Delhi. IGNOU.

Lancarter, F.W. (1968). Information Retrieval systems : Characteristics, Testing and Evaluation. N.Y. : Wiley.

Mooers, C.N (1963). The educational challenge of information science. In automation and scientific communication. 26th annual meeting of AD1, 1963.

Oddy, R.N. (1977). Information retrieval through man-machine dialogue. *J of Documentation*, 33; 1-4.

Rijsbergen, C.J. Van (1979) Information retrieval. 2nd ed. London: Butterworths, Reprint ed.-198).

Sparck Jones K (1971). Automatic keyword verification for information studies. London: Butterworths.

Vickery, B.C. (1970). Techniques of information retrieval. London: Butterworths.