UNITED ARAB REPUBLIC

THE INSTITUTE OF NATIONAL PLANNING

Memo No. 559

Information Retrieval; And Its Potential For Progress Of Research In The U.A.R.

By

Engineer Mohamed A. Mongy
Mrs. Aida M. Elewa
Abdel Hamid M. Dawash
Mahmoud I. Hendi

Supervised by
Dr. Ahmad Badr

Operations Research Center
April, 1965

Information Retrieval

And Its potential for Progress of Research In the U.A.R.

By

Engineer Mohamed A. Mongy

Mrs. Aida M. Elewa

Operations Research Center

and

Abdel Hamid M. Dawash

Mahmoud I. Hendi

National Information and Documentation Center

Programs prepared by: Mrs. Mary N. Youssef

Collator work by : Ramadan Abdel Mouti

Supervised by

Dr. Ahmad Badr

National Information and Documentation Center

March, 1965.

Acknowledgements

The group wishes to express his appreciation and gratitude to Dr. Salah Hamed , Director of the Operations Research Center and to Dr. Ahmed Kabesh, Director of the National Information and Documentation Center for their initiation . in calling for this joint study between the two institutions as well as for their enthusiasm and continued encouragement of the group.

Table of Contents

Pages

Introduction

This is the first report prepared jointly between the Operations Research Center, Institute of National Planning, and the National Information and Documentation Center, Ministry of Scientific Research, as a result of the effort carried out by the group.

The purpose of this study has two dimensions. The first is to acquaint and train a group of Scientists in the field of documentation and information retrieval about the present state of information and documentation techniques from the theoretical points of view as well as from the equally important applications, examplified in this first case with the equipment available at the Operations Research Center.

The second dimension is to stimulate a deeper consciousness among governmental, academic and industrial institutions about potentialities of information, information retrieval techniques and their important role in the development process and particularly in the progress of science and technology which is the bulwark of the national economy.

The joint cooperation between both institutions in the field of information retrieval is a pioneering and rewarding experience. This example should be continued between the two institutions and also should be followed among other institutions interested in the information retrieval field.

Ahmed Bedr

Chapter I

The Need For Information Retrieval

1.1. The Value of Information:

The value of information as an essential prerequisite for progress and development has become an undisputed fact in both advanced and developing countries. If countries have realized early the value of education in the process of development they have now mobilized a great part of their efforts towards the maximum accessibility of information because of its great role in developing their economies based principally on science and technology.

The extent to which the research worker depends upon the work of others has been clearly stated by one of the greatest of all scientists, the atomic physicist, Ernest Rutherford. As quoted by James Newmanin a recent issue of the "Scientific American", Lord Rutherford said:

"I have also tried to show you that it is not in the nature of things for any one man to make a sudden violent discovery; science goes step by step, and every man depends on the work of his predecessors. When you hear of a sudden unexpected discovery a bolt from the blue as it were - you can always be sure that it has grown up by the influence of man on another, and it is his mutual influence which makes the enormous possibility of scientific advance.

Scientists are not dependent on the ideas of a single man, but on the combined wisdom of thousands of men, all thinking the same problem, and each doing his little bit to add to the great structure of knowledge which is gradually being erected".(1)

As a commodity, information has two characteristics which can give it monitary value:

i. Relevence.

ii. Timeliness.

If either of these characteristics is missing, the information

(1) U.S., Congress, Senate, Committee on Government Operation, Science Program, Report No. 120, 86TH Congress, First Sess., 1959, p. 120.

is useless.

Information centers are operated in order to provide information to those who need it promptly and comprehensively from the world - wide literature.

The uses of such information are:-

a. It makes possible the completion of some projects which otherwise would have to be delayed or abandoned for lack of knowhow.

b. It saves valuable time for research and other decision - making personnel.

c. It helps develop knowledgeable and effective management at all levels of an organization.

d. It avoids undesireable duplication of research or other efforts and thus saves money which can be alloted to other productive purposes.

1.2. The Information "Problem" and the Information "Explosion":

Nowadays published and unpublished literature is increasing in quantity and complexity as a consequence of the expansion of human intellectual activities which resulted in a flood of literature and recorded knowledge unmatched in previous history.

It is estimated that about 55,000 periodicals are published annually containing about 1,200,000 articles of potential interest. Published literature includes also about 60,000 books as well as about 100,000 research reports and pamphlets..etc. Within this vast body of world wide research, published and unpublished, in different languages lie the information that research workers need to perform their work.(1)

More information has been produced than can be stored and retrieved. At the same time the need for effective exploitation of the literature is increasing because of the increasing

(١) أحمد بدر ، التعاون الدولى فى ميدان التوثيق العلمى ، مجلة المكتبة العربية : المجلد ١ عدد ١ ، ١٩٦٣ ، ٥٠ — ٥١

pressures for avoidance of repetition and duplication of research. This is a common problem in both developed and less developed countries. The problem may be more acute in the developing nations because they cannot afford to duplicate already done research.

Abilities of individual libraries, documentation centers & organizations to cope with this flood of literature are decreasing. The traditional library methods in storage & retrieval of this information are no longer adequate to respond to the complex demands and information requests of engineers, scientists, managers, decision makers... etc..

Thus the trend and the need to develop traditional library & documentation techniques and to use mechanical methods & devices is pressing in order to collect, store, organize, dessiminate and retrieve the information.

1.3. Traditional versus Non-Conventional Information Retrieval Systems:

Traditional library methods and techniques in analysis, organization and storage of information are valid only to meet limited requirements of research - workers in location and identification of needed information. But the complex and diversified needs and requests of research-workes nowadays have proven to be only met by advanced methods and techniques in which both the machine and the humanbeing constitute a functional information retrieval system.

It has to be asserted that the long developed and adopted library techniques & procedures especially cataloguing, classification, circulation as well as reference services constitute the basis from which different information retrieval systems have emanated as evolutionary (and may be revolutionary) to respond to the previous mentioned complex & diversified needs of research workers.

The greater degree of depth of analysis of information contained in documents is one of the main characteristics of non-conventional systems and which make them superior to traditional ones.

In a comparison between traditional and non-conventional systems, it can be said that analysis of information occurs by the following three methods:

Classification - Abstracting - Indexing.

1.3.1 Classification :

It is a systematic logical arrangement of index enteries usually in a hierarchical or tree pattern. The standard library classification systems, such as Dewey, Decimal, Bliss, Cutter, Library of Congress and Universal Decimal, all try to be hierarchical systems. The terms are arranged so that they proceed from the most general to the most specific.(1)

The traditional library classification (e.g. Dewey Decimal Classification or Library of Congress) is a rigid classification which is often called a "pigeonhole" classification and it involves the characterization of each record from a single point of view. When a record is to be stored, and only a single copy is available, a pigeonhole, or a single physical location, must be provided for this record.

Most classifications are artificial or synthetic and when these classifications are developed still represent, individually, a single rigid approach to a subject. This does not coincide with the needs and viewpoint of the searcher because records which contain information are generally multidimensional in nature. Hence the physical classification of records often tends to assume the characteristics of a rigid classification.

Information retrieval systems stress the utilization of multidimensional approach in classification in order to respond to the multidimentional requests of researchers and also in

(1) IBM, Reference Manual, Index Organization for Information Retrieval, p.9.

order to avoid the problems of the expansion of notation of classification.

1.3.2. Abstracting:

An abstract is a summary of a publication or article accompanied by an adeguate bibliographical description to enable the research worker to trace the publication or the article. Three types of abstracts can be identified:-

a. Traditional abstract: which may be either descriptive or informative. The former kind embodies a general statement of the scope of a document but this abstract does not substitute reading the original document. The informative abstract is written in order to provide a concise but comprehensive summary of the significant items of the document and this kind of abstracts may serve as a substitute for reading the original document.

b. Extract: and this is analogous to an abstract in that it represents what is considered by an analyst to be the important subject matter of a graphic record and this extract is selected from the original words of the documents.

Extracts may be selected by human analysts, or by the application of machine techniques to produce the "auto-abstract"(1)

The techniques used by humans to prepare extracts are subjective, and involve the exercise of judgement by an analyst in order to determine which portion of a document is of sufficient potential significance to warrant recording. When a machine is used for extracting, the entire text of a record is converted to machine - readable form.

---

(1) H.P. Luhn, "The automatic creation of Literature. Abstracts". IBM Journal of Research and Development, 2, No. 2 (April 1959), pp. 159 - 165.

It is then scanned by a digital computer. It is assumed when applying these methods that the frequency and distribution of keywords in the text can be used as the basis for determining the relative significant sentences in the text. Following this assumption, the sentences that are highest in "significance" (as determined by their high content of keywords) are printed out to produce an extract or "auto - abstract".

c. Telegraphic abstract:(1)

A telegraphic abstract is a detailed index to a graphic record. It is composed of:

(i) Significant words selected from the document,
(ii) Code symbols called role indicators which supply a context for the selected words, and
(iii) Punctuation symbols which separate and group the words and role indicators into various units in somewhat the same fashion as conventional punctuation does.

A file of telegraphic abstracts, though it is not used in the same way, serves the same purpose as a conventional index or card catalog - that is, to locate literature on a given subject.

The finished product of telegraphic abstract is a reel of tape with the information which the abstracters have partly furnished, translated into a computer code ready for searching by machine.

The purpose of the telegraphic abstracts, then is to provide "input" to the machine in a consistent and predictable form so that the machine can be programmed to search for certain predictable forms of information within this input. Details of application of telegraphic abstract is discussed in chapter (3) about the Western Reserve University information retrieval system.

(1) For definition of different types of abstracts see: kent, Allen, Text book on Mechanized Information Retrieval, pp.100-105.

1.3.3. Indexing:

Indexing involves the generation of index terms and the systematic ordering of these terms into an index. An index term is a device used to characterize the information content of a document.

The group of index terms used in the information system constitutes the list or dictionary of index terms.

Indexing may be performed in two different ways. One is to derive appropriate index terms (derived index) directly from incoming documents by lifting certain words or phrases from the text. The second way is to assign appropriate terms to documents from a pre-established index (assigned index).(1)

The form of indexing that is simplest to apply is the one that assumes on the part of the indexer the least amount of subject - matter background and the least amount of technical skill in indexing. It is in this type of indexing that a machine can perform with precision.

Indexing involves the traditional library techniques in subject analysis, namely the use of subject headings, which can be chosen from a subject authority list. In information retrieval systems-both manual and mechanical-the same principle of subject analysis is applied, but more depth of analysis is the main objective of these systems.

If we call the individual terms as subject headings in traditional methods we refer to these terms in nonconventional systems as descriptors, keywords, key-terms, discriminators, identifiers, or uniterms.

The two types of indexing can also be mentioned as follows:

A) Word indexing,

B) Controlled indexing

Word indexing, which involves the selection of words from a document, and their use as index entries. Whereas controlled

(1) IBM, general Information Manual; Storage, Retrieval and Dissemination of Information, p.6.

indexing implies a careful selection of terminology, for storage in the index, in order to avoid as far as possible, the scattering of related subjects under different headings. The control may be imposed by limiting the indexing in (a) the subjects that may be chosen.

(a) the subjects that may be chosen.
(b) the number of aspects that may be chosen,
(c) the language used to express the results of analysis.

Examples of two famous word indexing systems can be mentioned.

(i) Key - Word - In - Context (KWIC) or Permutation index:

This is an auto - encoding system where a computer recognizes significant words (keywords) and these words would be permuted to a predetermined position for alphabetization. When this process is applied to a title of an article or its abstract, all nonsignificant or "common" words are ignored. The following example is an illustration of a KWIC index prepared by IBM company about the titles of articles published on information retrieval and machine translation. (page 9)

(ii) Uniterm Indexing:

This is an index which involves the analysis of the contents of graphic records in terms of single descriptors called "uniterms" to define a document and to facilitate the manual coordination of these descriptors.

For example, using the following subject headings Zirconium - Physical properties - Tensile strength - High temperature, a card would be prepared for each descriptor used (term card) and the document numbers, referring to the document that contain this information, punched into these cards. When references are wanted covering this complex subject, the appropriate term cards are pulled and matched. All document numbers which appear on all four cards will contain information on the high temperature strength of zirconium. If one is searching for the more general topic of physical properties of Zirconium then a match of two term cards Zirconium and Physical Properties will also retrieve these documents.