Chapter 1

Introduction to Information Retrieval Systems

INFORMATION STORAGE AND RETRIEVAL

THE FIELD

The major determinants behind current information storage and retrieval efforts are the great volume of data pouring from our printing presses and our inability to locate much of it after it has appeared.

Responsibility for storage and retrieval of printed information has traditionally rested with the librarian. Early libraries concentrated on arranging books in some prescribed order on shelves. As the number of books increased, a complex organisation became necessary in order to make the contents of a library collection more readily accessible. To provide such organisation, librarians developed subject-classification schemes, the card catalogue, and other tools. These bibliographic devices now constitute the basic structure for control of library collections and are the fundamental finding aids that researchers employ.

Although conventional library tools today make location of a particular title among miles and miles of shelving a routine and simple task, they are not designed to provide more than a rough-cut approach to the subjects covered. For the users of general libraries this may be all that is needed, but when the same subject-classification techniques are applied to highly specialized collections of non-book, technically detailed data, the imprecision of such methods of content retrieval becomes apparent. Because all knowledge and language are dynamic, constantly changing processes, any subject classification becomes obsolete almost from the moment of its creation. Furthermore, as one moves into increasingly specialized areas of knowledge, research becomes more complex. As new ideas generate new facts and new terminology, the task of organising them and establishing their proper relationship to one another becomes ever more difficult.

An important distinction has been made between systems that locate documents and systems that produce information. Yehoshua Bar-Hillel has emphasized the difference between "literature searching" and "information retrieval," pointing out that the problems of storing and retrieving documents should be considered apart from problems concerning information. Literature searching, he contends, involves determining which documents or books are relevant to a chosen topic. Information retrieval is the act of obtaining answers to questions about a selected subject (1957).

Emphasis has thus been placed on finding new ways and means of codifying or indexing data so that they will lend themselves to correlation at time of searching. The trend has been to achieve greater depth of content analysis. Not only have new analytical methods been devised but investigations have also been made of the feasibility of employing electronic machines for analysis, storage, and retrieval of information. Because of the great mass of data involved, new storage and handling techniques may have to be invented; these techniques must be more advanced than those customarily used for manually shelving books and filing documents. The emergence of information storage and retrieval as a new field reflects an awareness among librarians and others that the selection and manipulation of fragments of information, rather than of entire documents, will require unconventional tools.

A critical need for more advanced information systems has evolved because of the steady growth of publishing and the complex ways in which information has come in recent years to pervade decision-making processes in business, science, and government. References to the effects of expanded publishing were made by Fremont Rider (1944) and Vannevar Bush (1945). Shortly thereafter, the implications of the "information explosion" in science and technology were discussed at the first international conference on the subject, held in London by the Royal Society (1948). At that time, it was already clear that the publishing rate in science and technology was increasing exponentially and that specialization in individual sciences and the development of interdisciplinary research were generating multiple uses for the same information. Although interest in information storage and retrieval thus received its start in the world of science, it soon spread to other areas, particularly business, industry, and government and, notably, to institutions like the U.S. Library of Congress.

Another factor responsible for the independent development of the field of information storage and retrieval has been the impact of technology. Research and development in the computer sciences, the photographic industry, and signal communication promise to provide powerful new methods and techniques for information storage and retrieval. Modern data-processing equipment has already been successfully applied to the numerical areas of scientific computation and business operations. The prospect of being able to use computers to solve non-numerical problems that involve natural language has been a major impetus encouraging the evolution of advanced information storage and retrieval techniques. The appearance in 1948 of Shannon's theoretical foundation for a general theory of information stimulated researchers to investigate the possibility of applying the principles of mathematics to the problems of information communication, by means of computers.

The field of information storage and retrieval involves librarians, documentalists, mathematicians, system designers, linguists, equipment manufacturers, operations researchers, and computer programmers, among others. All are concerned with methods of expediting the prompt retrieval of information in such diverse areas as libraries, business and industry, military command and control, and scientific research. Because the field is interdisciplinary, considerable confusion regarding the boundaries of the effort has existed. A comprehensive bibliography covering the broad spectrum of interest appeared in 1958, and an introductory textbock on the subject was published in 1963.

Classification of Subjects in Documents

Several specialists have devoted themselves to research into the problems of information organisation. Among them is Mortimer Taube, who is identified with the concept of *coordinate indexing,* which provides a method of coordinating index terms as combinations rather than permutations. Taube called his index terms*uniterms,* and a coordinate index consists of a set of uniterm cards on which appear the identification numbers of the documents relevant to each uniterm. Searching is accomplished by selecting those uniterm cards pertinent to a request and correlating their document numbers. Matching numbers represent those documents for which the uniterms are simultaneously relevant.

Calvin Mooers, one of the earliest proponents of coordinate indexing, proposed a concept of storing in one fixed place the codes for the subjects in a document, one code being *superimposed* on another. This technique is particularly applicable where coding space is at a premium, such as on edge-notched cards. Mooers also conducted extensive research into the mathematical structure of coding.

James W. Perry and Allen Kent have advanced the idea of the so-called *telegraphic abstract,* in which a*phrase* represents the logical unit of thought in a document, *subphrases* represent the individual words and concepts, and *role-indicators* describe the role that a particular word plays in the *phrase.* By using this method it is possible to describe a document in an artificial language or system that has more meaning than the sum of separately assigned subjects. *Faceted classification,* still another technique for organising concepts expressed in documents, has been examined by S. R. Ranganathan and Brian C. Vickery.

Computer Analysis of Natural Language

A number of experiments have been conducted, and corresponding computer programmes have been written, on the possibility of using computers to perform quasi-intellectual functions. Within the past few years, increasing emphasis has been placed on machine analysis of the syntax and semantics of natural language. This has led to the development of computer programmes for such functions as language data-processing, machine translation, automatic indexing, automatic abstracting, concord ance building, and text condensation. Still other computer programmes have been written for the preparation of permuted title indexes as well as conventional printed indexes. Several researchers have produced computer programmes that embody sophisticated mathematical principles for searching natural language. Research work has explored ways of extracting meaning from a text by means of word association, syntactical analysis, and even contextual analysis. M. E. Maron and J. L. Kuhns have applied the calculus of probability to automatic indexing in an attempt to establish a theory of relevance.

Converting Text to Machine Readable Form

The ability to convert original data automatically from the printed page to an input form usable by machines is fundamental if electronic computers are to be employed in work involving information. Until this becomes possible the use of computers cannot be considered economical. In the absence of automatic conversion equipment it is necessary either to type or keypunch the data over again. These processes are expensive, slow, and unreliable. For these reasons, efforts are continuing to produce *character-recognition* machines. These are devices engineered to scan automatically the letters, words, and sentences of a text and to convert them directly into discrete digital representations.

The goal is to "read" rapidly large quantities of printed information, so that further processing of the data can be performed by a computer. Optical scanning and magnetic-ink reading are the two most common character recognition techniques in use. Thus far, only alphanumeric data in a prescribed type font are readable by machine. Research in auditory recognition is also under way to determine whether a machine can automatically discriminate phonetic sounds and, in so doing, produce a satisfactory digital code for input to a computer.

COMPACT STORAGE OF SOURCE MATERIAL

Microfilm is, at present, the most effective means of storing original documents and of thereby controlling their volume. An impressive array of different cameras and a multiplicity of microfilm media are available commercially. Roll film, aperture cards, film in cartridges, microfiche, sheet film, and microcards are but a few of the examples of common microforms at present in use in information installations.

New dry processes have been introduced to overcome the disadvantages of wet chemical development, which is normally associated with the silver-halide film process. Diazo, for example, is a film which is exposed with ultraviolet light and developed in gaseous ammonia. Kalvar, another film, is exposed with ultraviolet light but is developed with heat at a temperature equivalent to that of a warm iron. The latest dry process is photochromies, which claims data-compression ratios of up to 400:1 with practically no loss of resolution. Photochromic film is exposed with ultraviolet light and can be erased, if necessary, with white light.

Printed material that is compressed into a microform calls for auxiliary equipment—inspection viewers, service viewers, and printing equipment for individual page copying. Equipment available on the market makes it possible to view any microform and to obtain a copy of an entire page or part of a page in seconds. Devices that fall into this category provide push-button copying, frame by frame, using manual, semiautomatic, or fully automatic auxiliary means.

Ever since Vannevar Bush proposed a Memex machine in 1945, much equipment has been designed to combine the dense-storage capability of film with the searching speed of electronics. The Rapid Selector, the first such device to be built, recorded frames of abstracts and corresponding digital codes on a 2,000-foot reel of 35-mm. film.

Following the Rapid Selector, other equipment appeared, such as Minicard, Media, Flip, File-search, Lodestar, Verac, and Walnut. Originally, each of these devices was designed to satisfy the needs of highly specialized information-system customers, but all of them represent technical progress towards combined use of electronic and photographic media for many purposes.

Minicard and Media are systems which store digital and graphic information on chips of film. Flip, Filesearch, and Lodestar, on the other hand, require the stored information to be contained in sequence on reels of film. Verac and Walnut are slightly different. The former uses a store of glass plates on which a matrix of images is recorded; the latter, strips of Diazo film for recording. Both, like all the others, have electronic-searching capabilities.

Communication of Information

No discussion of the technologies pertinent to the field of information storage and retrieval would be complete without consideration of the role of communication.

In the early 1950s, RCA conducted a demonstration of Ultrafax at the Library of Congress. A film copy of*Gone With the Wind* was sent over communication lines to a receiving point in a distant city. This facsimile transmission heralded the use of communication facilities for the transfer of visual data from one point to another. Video recording and transmission provide still another medium for sending graphic information over great distances.

Retrieval at a distance of digital and graphic-information presupposes the availability of an interconnected communications network. On this assumption, research has been conducted to explore the relationship between man and machine in order to define more clearly the division of tasks between them. This in turn has led to further research of on-line systems, which establish direct communication between the man at an input-output console and the computer. The Massachusetts Institute of Technology has led the research