INFORMATION RETRIEVAL
Computational and Theoretical Aspects

H. S. HEAPS
Department of Computer Science
Concordia University
Montreal, Quebec
Canada

ACADEMIC PRESS Ne
A Subsidiary of Harcourt Brac

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data

Heaps, H S
Information retrieval, computational and theoretical aspects.

(Library and information science)
Includes bibliographical references.
1. Information storage and retrieval systems.
I. Title. II. Series.
Z699.H38 029.7 78-3338
ISBN 0-12-335750-0

PRINTED IN THE UNITED STATES OF AMERICA

Contents

1

Introduction

1.1 Growth of Recorded Knowledge

Information retrieval is a new discipline in the sense that many of its modern applications depend on concepts that have been formulated only during the past few decades. Nevertheless the subject has roots that extend back through many centuries.

The traditional library, as a collection of documents, led to the development of standard procedures for manual cataloguing, use of card indexes, bibliographies, and the circulation and ordering of books, journals, and reports. However, the traditional library was oriented more to the provision of documents than to the supply of information. This orientation is efficient provided that library users are interested primarily in well-defined subjects covered by a small number of books and journals. Yet at the present time there are many fields of study whose investigation requires information from a number of different disciplines, and for which requests for relevant information cannot be met by reference to a small, easily specified set of documents.

As remarked by C. P. Snow,[1] during all of human history until the present century the rate of social change has been sufficiently slow as to

[1] C. P. Snow, *The Two Cultures and the Scientific Revolution* (Cambridge: Cambridge University Press, 1959), p. 45.

pass unnoticed in one person's lifetime. This is no longer so. People are generally aware that social and technological changes are taking place very rapidly and are the result of discoveries that are understood by only a few specialists. Yet the consequences of many new discoveries affect the lives of entire populations to a degree that has never been the case previously. Thus more and more people are vitally interested in having fast access to more and more information.

The ability to maintain the rapid growth of technological and social changes requires that vast amounts of information be instantly available when required. The problem involved in meeting such a requirement may be emphasized by noting that it has been estimated that the number of scientific journals that existed in the year 1800, 1850, 1900, and 1966 was approximately 100, 1000, 10,000, and 100,000, respectively.[2] Holt and Schrank[3] have indicated that between 1920 and 1960 the periodical literature in economics increased from 5000 to 40,000 articles per year. Likewise, the periodical literature in psychology increased from 30,000 to 90,000 articles per year. The annual number of papers published in mathematics during the period from 1868 to 1966 increased from about 800 to 13,000.[4,5]

For a number of different disciplines the increases in the number of articles published in periodicals over the period between 1960 and 1970 have been estimated by Carter as shown in Table 1.1.[6] Presumably the disagreement with Holt and Schrank's estimate for psychology is caused by a different definition of what constitutes a periodical article in psychology. The particular values of the individual figures are, however, less important than the implied growth rates of values estimated according to the same criteria.

The manner in which a new discovery may lead to an increase in the number of publications may be illustrated by noting the growth caused by the proposal of the laser structure in 1958.[6] In 1960 there were approximately 20 papers on the subject of the ruby laser; in 1961 there were approximately 100 on the helium–neon laser; in 1962 there were 325 on the solid-state laser; in 1963 there were 700 papers on the GaAs laser,

[2] Proceedings of the Royal Institute of Great Britain, vol. 41, Part I, 1966.

[3] C. C. Holt and W. E. Schrank, "Growth of the professional literature in economics and other fields, and some implications," *American Documentation* 19(1968):18–26.

[4] K. O. May, "Quantitative Growth of the mathematical literature," *Science* 154 (1966):1672–1673.

[5] K. O. May, "Growth and quality in the mathematical literature," ISIS 59(1968):363–371.

[6] A. Neelameghan, "Theoretical Foundation for UDC: Its need and formulation," *Proceedings of the International Symposium*, Herceg Novi, Yugoslavia, June 28–July 1, 1971.

Table 1.1
Number of Articles Published in Periodicals During 1960 and 1970.

Subject	1960	1970
Mathematics	15,000	30,000
Physics	75,000	155,000
Civil engineering	15,000	15,000
Mechanical engineering	10,000	20,000
Electrical and electronic engineering	80,000	150,000
Aerospace engineering	35,000	75,000
Industrial engineering	15,000	15,000
Chemistry	150,000	260,000
Metallurgy	35,000	50,000
Biology	150,000	290,000
Geosciences	91,000	158,000
Agriculture	150,000	260,000
Medicine	220,000	390,000
Psychology	15,000	30,000
Other subjects	929,000	1,882,000
Totals	1,985,000	3,780,000

pulsed laser, and *Q* switching; in 1964 there were 1000 on the ion laser; and in 1965 there were 1200 papers on the N_2–CO_2 high efficiency laser.

The term *information explosion* has been accepted very readily by workers in scientific fields who tend to be extremely conscious of the possibility of being unaware of work done previously, or being undertaken concurrently, by other scientists. There has even been a tendency, perhaps, to exaggerate the consequences of overlooking the work of others.[7]

The information needs of physicists and chemists engaged in research, administration, and teaching have been discussed in the literature.[8] The roles of abstracting services in physics and summary papers in chemistry have been discussed respectively by Urquart[9] and Bernal.[10] A general discussion of the sources of information available to chemists and physi-

[7] A. G. Oettinger, "An essay in information retrieval or the birth of a myth," *Information and Control* 8 (1965):64–79.

[8] Survey of information needs of physicists and chemists. The report of a survey undertaken in 1963–4, in association with Professor B. H. Flowers, on behalf of the Advisory Council on Scientific Policy. Journal of Documentation, vol. 21, pp. 83–112, 1965.

[9] D. J. Urquart, "Physics abstracting use and users," *Journal of Documentation* 21(1965):113–121.

[10] J. D. Bernal, "Summary papers and summary journals in chemistry," *Journal of Documentation* 21(1965):122–127.

cists was given by Bottle.[11] The need for ready access to literature in the social sciences has been discussed by Guttsman.[12]

1.2 The Discipline of Information Retrieval

Although recorded information usually is retrieved by means of stored data that represents documents, it is the emphasis on information relevant to a request, rather than direct specification of a document, that characterizes the modern subject of information retrieval. In addition to being concerned with the practical aspects of the design of operational computerized retrieval systems the subject of information retrieval includes aspects of the theory of measurement and definition, and of information content and relevance.

Some aspects of information retrieval may be compared to statistical communication theory as applied by electrical engineers and applied mathematicians during the past 30 years. The problem of locating relevant information from a body of widely dispersed knowledge is analogous to detection of the presence of a signal pulse in the presence of a noise background. Concepts such as the Wiener root-mean-square criterion, matched filters, feedback, and correlation detectors have their counterparts in the theory of information retrieval. It is perhaps rather curious that Norbert Wiener, who displayed great insight in the application of linear prediction techniques to problems in control theory and cybernetics, was very sceptical of the value of studies in information retrieval as he claimed that any information that he might require for study of his own subject could be obtained most easily by writing to any of the half dozen world experts in the field.

Wiener's point of view is interesting in that it serves to emphasize the difference between the environments of today's scholars and those of previous generations. Just as the number of people presently alive is a surprisingly large proportion of all people who have ever been born, so the amount of information in recorded form is many times larger than at any previous time. Moreover, not only is scientific, technological, and sociological information required by experts in these fields but also by nonspecialists who are unlikely to know the names of the leading authorities.

The rate of increase of available information has many philosophical

[11] R. T. Bottle, "A user's assessment of current awareness services," *Journal of Documentation* 21(1965): 177–189.

[12] W. L. Guttsman, "The literature of the social sciences and provision for research in them," *Journal of Documentation* 22(1966): 186–194.

implications. The concept of man as a solitary traveller through time with some interaction from other human beings, and later the concept of man surrounded by a mechanistic universe, may now be replaced by the idea of man as an information receiver. Since information does not necessarily relate to physical quantities or to precisely measurable terms, it might be speculated that techniques developed for information retrieval and information evaluation eventually should be developed in a direction leading to further understanding of the process by which human beings associate ideas and gain understanding of scientific and humanistic concepts.

Spoken communication between humans is by sound waves that travel a distance of about 1000 feet in 1 sec. Communication between computers, or between computers and peripheral devices, is by electrical impulses that travel approximately 1000 feet in 1 μsec. For a simple introduction to the principles of data transmission between computers, or through communication networks, the reader may refer to a paper of Kallenbach.[13]

Relays and switching devices within modern computers have response times of a few nanoseconds (1 nsec = 10^{-9} sec). Consequently, large amounts of information may be processed and transmitted by computers in a short time interval. Scientifically, well-informed, imaginative speculation on the possible consequences of utilizing such high transmission rates, and the ability of computers to absorb information at such speed, has led to many science-fiction writings such as, for example, those of the astronomer Hoyle.[14-16] The degree to which instant communication has already affected contemporary society has been particularly emphasized from a popular viewpoint by McLuhan.[17]

Computerized information retrieval systems must be economical as well as feasible. In the same manner that economic considerations have led to more critical and hence more mathematically sophisticated design of engineering structures and chemical engineering processes, it is the economic considerations that are leading to a requirement for more precise mathematical formulation of the principles of information retrieval in order to ensure that the computers and computer accessible storage devices are used in an economic manner. Concurrently with the develop-

[13] P. A. Kallenbach, "Introduction to data transmission for information retrieval," *Information Processing and Management* 11(1975):137–145.

[14] F. Hoyle and G. Hoyle, *A for Andromeda* (Greenwich, Connecticut: Fawcett Publications Inc.).

[15] F. Hoyle, *The Black Cloud* (London: Heinemann, 1957).

[16] F. Hoyle, *October the First Is Too Late* (London: Heinemann, 1966).

[17] M. McLuhan, *Understanding Media: The Extensions of Man* (New York: McGraw-Hill, 1964).

ment of new theory it is also important to study the behavior of operational retrieval systems in order to gain insight into the problems that arise from the point of view of the users of the system.

To some extent, the efficiency of a computerized information retrieval system is dependent on the computer hardware. Since computer hardware is continually being improved, and since computers become obsolete with the advent of more powerful and more economical ones, it is likely that operational retrieval systems will continue to be subject to constant revision and expansion. This fact presents less difficulty than might first be thought since it is the hardware and computer programs that change; the vast quantities of stored information may remain unchanged or be converted within the computer.

In analysis of complex or vaguely defined problems, the importance of good notation cannot be overemphasized. An efficient notation allows concepts to be stated clearly and allows data to be visualized as a whole, instead of as a number of isolated items. It also enables problems to be formulated in a precise manner with a true appreciation of any assumptions involved.

For formulation of information retrieval problems the concept of matrix transformation often allows statements to be made in a very compact form. The structure of written text, including the extent to which it is predictable, may be analyzed by probability theory.

Information theory, as developed in mathematical terms by Shannon,[18] is concerned with information content in terms of the amount of information required for identification of symbols or words rather than in terms of the knowledge communicated by them. Randomness, or uncertainty or lack of knowledge, is measured in terms of entropy. The power and limitations of the theory have been discussed by a number of authors.[19] Extension of Shannon's theory and computer analysis of written text to analyze information content, subject matter, and style, offers a number of problems to the student of information science. It also suggests the possibility of further cooperation between computer scientists and workers in other disciplines.

A study of information retrieval is necessarily concerned with optimization since it is desired to retrieve relevant items in the shortest possible time, or with minimum expense, or with maximum efficiency in regard to some estimate of relevance. The measure of relevance may be formulated

[18] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal* 27(1948):379–423, 623–656.

[19] P. L. Garvin (ed.), *Natural Language and the Computer* (New York: McGraw-Hill, 1963).

in mathematical terms and leads to considerations based on the mathematical theory of pattern recognition.

The subject of information retrieval is thus developing through application of matrix notation, probability theory, optimization techniques, pattern recognition, and systems analysis through which operations are represented by mathematical models that may be programmed on a computer.

The practical problems of information retrieval involve sufficient quantities of data that the introduction of any reasonable attempt to be systematic and independent of human intuition involves the handling of a large amount of data. The greater the degree of sophistication required, the greater the task of dealing with the information, and the greater the need for computer use since modern computers can store large quantities of data on computer accessible files and can examine the data very rapidly.

The systems analyst who designs a computer program for information retrieval or for some aspect of library automation first represents the entire operation by a mathematical model. The subsequent computer programming then proceeds in relation to this assumed model, which is described to the computer by means of a program written in some computer language. The resulting situation is that when the computerized system is in use, the librarians who use it tend to describe and evaluate it in terms of traditional library concepts while the computer analysts and programmers describe it using a different language and tend to think about the mathematical model rather than the operational system. It is clearly necessary for the two groups of people, those with traditional library backgrounds and those with computer science backgrounds, to have sufficient knowledge of each others' problems and the means of describing them. Otherwise, no effective communication may occur between the two groups. Ability to treat the problems from a wider viewpoint than generally acquired in the course of either a library science or computer science education is one of the skills expected of the information scientist.

The purpose of the present text is twofold. On the one hand an attempt is made to introduce the computer science student to some of the basic problems of information retrieval and to describe the techniques required to develop suitable computer programs. On the other hand an attempt is made to describe the general structure of the relevant computer programs so that basic design considerations may be understood by information officers and librarians not well versed in the details of computer science.

The ability of computers to perform complex arithmetic operations is more important for scientific computations than for information retrieval. The problems that arise in information retrieval are more concerned with identification, storage, rearrangement, and sorting of alphabetic data. It is the ability of computers to be used for these tasks that makes them of use in information retrieval.

Efficiency of a retrieval system is sometimes defined solely in terms of user satisfaction with the items retrieved. If efficiency is also related to economics then the search procedure should produce the desired results through utilization of minimum computer time and with compact storage of the information in computer memory or in computer accessible files. It is most important that all who are concerned with the system design should appreciate the fact that user convenience and economic efficiency are not always compatible. Both the computer scientist and the non-computer-oriented information specialist should be aware of the price that may have to be paid for an increased emphasis on either factor.

1.3 Computer Learning and Adaptive Systems

One measure of the efficiency of an information retrieval system, whether manual or computerized, is in terms of the user's estimate of the relevance of the documents retrieved and the nonrelevance of the documents not retrieved by the system. If the retrieval procedure is dependent on parameters that may be varied within the system, then in response to certain requests the parameters may be varied until a maximum efficiency is believed to result. The same values of the parameters then may be assigned automatically whenever similar requests are encountered. The resulting system then may be said to have a learning capacity since in response to a given situation it behaves in a manner that was found to be efficient in dealing with similar situations in the past.

Retrieval of information by human beings is often dependent on the human ability to infer associations between different concepts and to generalize specific requests in order to cover broader fields of interests. Attempts have been made to formulate means by which computers may infer associations between different concepts. Such formulations usually are based on the premise that if one concept is often associated with a second one, and also is often associated with a third one, then there is a strong probability of some connection between the second and third concept even in the absence of the first. Clearly such a premise is not always true but is, nevertheless, true in many instances. Associations of such nature may be studied by means of term connection matrices as introduced in Chapter 10.

Since a computer has memory and is able to recognize and compare words, it may be programmed to recognize that certain words are equivalent in meaning, that certain words are broader or narrower in meaning than others, and that certain words have special meanings when used in association with other words. Thus a considerable amount of information based on human judgment and experience may be incorporated into a purely mechanistic system. Various aspects of this are discussed further in Chapters 11 and 12.

The basic problem in the design of systems with learning capacity is not one of computer programming, but rather is the formulation of what is meant by most efficient, by relevance, and by other terms of judgment. The problem is similar to computer recognition of patterns and is discussed in reference to automatic classification of documents in Chapter 13. Examination of the problems in more generality would lead to a study of topics in the branch of computer science known as *artificial intelligence*. This subject is concerned with the study of the use of machines to undertake tasks which, if performed by humans, would be said to indicate the exercise of intelligence.

1.4 Computer Identification of Meaning

Although we may readily appreciate that a computer may be programmed to compare, recognize, and identify words, it might be contended that a computer lacks human intuition and therefore cannot be expected to discover the meaning that a passage of text is able to convey to a human reader. However, the human being is conditioned by years of education, through learning by success and failure, by association of ideas, and by experience in evaluation of the consequences of assumptions.

When a body of text is read by a human it may cause the reader to respond by expressing an opinion, proving a theorem, ordering a book, or writing a literary criticism of the material read. Suppose the same text read into a computer memory causes the computer to print a statement, write an order for a book, write a literary criticism, or write a poem. This is clearly possible if the action may be made to depend on identification of textual words rather than on an understanding of them. Some aspects of information theory, based on identification rather than semantic meaning of text, are introduced in Chapter 8.

Identification of the subject matter of text is a problem that is suitable for present-day computers. It is of importance in computer preparation of abstracts, in computer selection of indexing terms, and in the use of computers to recognize the authorship of manuscripts of disputed origin.

Because computers may process large amounts of information in a short time, they may be used systematically to examine language for structural qualities not easily deduced by intuitive means. Although a computer has no understanding of the meaning of language it may be used to deduce structural relations of value in linguistics and in studies of relationships between different languages. It may be remarked, however, that the problem of machine translation of one language into another is a very difficult one and is significantly more complex than many of the problems that arise in information retrieval.

It appears that the languages of related disciplines form a system of intersecting vocabularies, each with its own characteristic structure. Each major discipline has its own basic vocabulary. For each subdiscipline there is a special vocabulary of the names of the things or concepts studied, and there is also a further well-defined vocabulary of the specialized terms that form the so-called jargon of the subdiscipline. The jargon words are not necessarily unique to the subdiscipline but have their special meanings only when used with the special vocabulary of the subdiscipline. Determination of these special vocabularies and their structures requires use of a large computer and is of considerable importance for future application of computers to analyze the meaning of printed text.

1.5 Document Retrieval, Library Automation, and Privacy of Files

Methods for search of computer accessible data bases that contain document references are discussed in Chapter 6. As the files grow in size, the problem of their storage in an economic manner becomes very important. It becomes necessary to consider various methods of compression so that storage costs may be reduced. Various means have been proposed and some are described in Chapter 9.

Computers may be used to automate many aspects of library operation, including circulation, cataloguing, ordering, and searching for books by specified authors or on particular topics. For library applications the computer is particularly useful if it may be accessed through remote terminals situated in the different departments of the library. All records may then be on magnetic tape or disk files at a central location, but may be searched or updated through remote terminals. If the computer has time-sharing capabilities it may work on other processing while at the same time appearing to be always available for library use. Library automation is not considered separately in the present text but constitutes an important practical application for many of the techniques discussed. It is particularly suitable for application of the compression techniques discussed in Chapter 9.