Computers in the Library

by MICHAEL F. LYNCH

Postgraduate School of Librarianship,
University of Sheffield

The past few years have seen many experiments in the automatic handling of information. It is too soon to know where they will lead

A NUMBER of novel information services are currently becoming available in Britain to workers in science, medicine and engineering. They include the MEDLARS service of the U.S. National Library of Medicine[1], provided jointly by the National Lending Library for Science and Technology and the Computing Laboratory of the University of Newcastle upon Tyne, the service based on *Chemical Titles*[2] and *Chemical–Biological Activities*[3,4] to be provided by the Chemical Society Research Unit in Dissemination and Retrieval of Information at the University of Nottingham, and that of the National Electronics Research Council[5]. The provision of these services is supported by the Office for Scientific and Technical Information of the Department of Education and Science, which also encourages research by means of grants and contracts. In these services, computers are used to select references to articles from the current periodical literature. The role of the computer in this operation, and the place of the new services in the British information system, are matters of current interest.

The application of computers to the selection of references in response to a request for information is not new. Almost 10 years ago, H. P. Luhn of the IBM Corporation experimented with selective dissemination of information by matching descriptions of scientists' interests with index terms applied to articles[6]. What has followed was stimulated by Luhn's work, but he is perhaps most widely known today for his KWIC—or Key-Word-In-Context—Index, a machine-produced concordance of the titles of articles produced by machine[7].

The success of the KWIC and other machine-produced title indexes (Chemical Abstracts Service launched *Chemical Titles* in 1961, and this has been followed by many similar publications) and the slower inception of computer based information services highlight the problems of retrieval. The computer is essentially a manipulator of symbols and not an information processing device in the strict sense. The success achieved in numerical computation and allied fields derives from the existence in these areas of meaningful rules or algorithms for processing symbols into some useful final state, the rules being incorporated into the computer programme. The relatively slow progress in applying computers to the field of scientific information can be ascribed to the inadequacy of rules for manipulating the symbols by which the information is recorded. Almost the sole established rule of use, when Luhn began his work, was the ordering of words in alphabetical sequence (although, as Metcalfe[8] has commented, this is one of our best precision tools, hence the success of the KWIC index). To use computers effectively for retrieval, we must first discover the appropriate rules.

The essence of the problem is that the relation between concepts and the words in which they are described is as yet imperfectly understood. For retrieval, this means that a request for information, phrased in a certain set of words, may be satisfied by an article, the words or the description of which may be an entirely different set. Much of the research in information retrieval (or, more precisely, reference retrieval) by computer has therefore been concerned with establishing computable models which bring these two sets of words into approximate coincidence, thus leading to the correct selection of references. The assessment of "correct" in this context is itself a most difficult task, as it depends primarily on human factors and, in the final analysis, on a judgment of usefulness. Much of the pioneering work in the evaluation of retrieval systems has been carried out by Cleverdon at Cranfield[9].

The devices traditionally used to overcome the problems of variant word descriptions are indexing and classification, and these retain their place in computer studies, although earlier workers tended to discount the importance of classification. Indexing is the selection or assignment of words describing the content of a document. It involves decisions that relate to the body of the documents being indexed, and to the questions likely to be put to the collection. The index words must both describe and discriminate. The indexing may be derivative (word indexing), using only words used in the document, or it may be by assignment (subject indexing), that is, words other than those used in the document may also be used. The former method is more amenable to mechanization, the latter to treatment by human indexers. The role of the representation, that is, the set of index words applied to the document, is critical. If a concept is not indexed, it cannot be retrieved directly, while the choice of words obviously affects the probability of retrieval in a given situation.

Class Distinction

Classification, in its widest sense, has to do with forms of relatedness, and with the organization and display of the relations in a useful manner. In classifying a set of items both the similarities of the items and the relations of the classes to one another must be taken into account. Some basis for judging similarity must be established. A library classification, the form in which it is most generally met, has as its main purpose to provide a linear arrangement of books so that those standing close together have most in common in regard to subject matter. Certain relations alone, those of a generic–specific and co-ordinate nature, can be accommodated, and many potentially useful relations must be excluded, or displayed by a limited number of cross-references. The extension of an existing classification and changes in its basic form are slow and costly processes. Yet classifications should be dynamic to allow for changes, especially in rapidly

developing fields. Even more important, they should accommodate all potentially useful forms of relatedness, and allow discrimination among them.

Indexing itself has an element of classification, more marked, certainly, in the case of subject indexing than in word indexing; documents for which the same term has been used can be presumed to have some similarity. Alphabetic arrangement of index words also tends to group like items together, as with "computation", "computer" and "computing". But what of "automation", "mechanization", etc.? Subject indexes overcome in part this difficulty by using a controlled vocabulary, in which a consistent choice is made among synonyms and closely related expressions. Thus "See" references guide the user of an index from the excluded term to the chosen one, and "See also" references include some, at least, of the associated concepts. Control of vocabulary, however, implies approximation (as also does condensation) and thus loss of information. None the less, a considerable degree of control of vocabulary is desirable in an index for human use, as it cuts down the number of points at which the user must consult the display.

Automatic indexing and automatic classification have both been extensively studied in connexion with machine retrieval. Luhn, again, was one of the first to use a computer to extract descriptive and discriminant words from the machine-readable text of a document by means of a statistical approach[10]. He linked this closely with automatic abstracting (more correctly, extraction), in that he sought first to determine the significant words of a text, and then to find a small number of sentences in which the words appeared most frequently, using the sentences as an abstract for the document. At about the same time, Baxendale included simple syntactic techniques to derive simple phrases, as well as single words, as index terms[11]. Luhn also studied means of controlling and correlating words in a vocabulary by compiling a thesaurus. Later work in this area, for example, the studies of Maron[12,13], has been concerned with refining the statistical criteria used in selecting the index terms. For example, the frequencies of words within the collection, as well as within a particular document, have been taken into account. Word indexing is, however, prone to excessive scatter because of synonyms and to the inclusion of irrelevant entries. Automatic methods for assignment indexing to reduce this scatter have also been investigated.

Mechanical Classification

Complementary to the investigations of automatic indexing has been the work on automatic classification. It is clear that if computers are to be used to retrieve references on the basis of records derived from them, classifications of word usage, rather than of concepts, must be developed. Thus Doyle[14] suggested the "semantic road map", a classification compiled automatically and derived from usage, which would help to guide searchers in the choice of appropriate terms. Later, Stiles[15] introduced the idea of the association factor in indexing. This is a measure of the frequency of co-occurrence of pairs of index terms in relation to the frequency of occurrence of each term. For each index term used in a collection, he generated a list of other terms which co-occurred with it to a significant extent. He termed this list the profile of each term. He was then able to detect synonyms and near synonyms in the vocabulary by comparing all the profiles and taking those which had more than a certain number of terms in common. The results of this and similar techniques, developed by Needham[16], Meetham[17], Vaswani[18] and others, are multi-dimensional classifications of words in the form of networks in which the nodes are words and the links among them are associations derived from usage. Indeed, an electrical network to represent the relations of such a classification was constructed by Giuliano[19], but they are more commonly represented in the form of a matrix within the computer. The links within the network are used to find words which are potentially associated with those of an enquiry. The word "water" might, for example, be found to be linked closely with "moisture", "humidity" and "dampness". Thus a match between the words of the query and those of a relevant document becomes more probable.

Increasing attention has recently been paid to the role of structure, both in the description of a document (as syntactic structure) and in classification, and elegant graphical theoretical techniques have been developed to deal with structural manipulations. Salton[20], Gardin[21] and Meetham[22] have contributed especially to these aspects.

System Effectiveness

Few of these techniques have yet been incorporated in generally available services. The adequacy of the models on which they are based can be judged only in terms of the usefulness of the results, and it is the difficulty of assessing this usefulness which limits the rate of progress towards fully automatic retrieval of references. The high cost of many of the sophisticated procedures is another important factor; in the first place, some of them require a record of the texts of documents which can be read by machine. Second, the computations involved are time-consuming and expensive. Two factors which work towards the reduction of these costs are the increasing power of computers in relation to their cost, and the trend towards computer typesetting in publication, which gives a digital record of the text as a by-product. One result of the latter factor is that the economics of searching the complete texts of documents, rather than sets of index terms, become more favourable. It is likely to be true for some years to come, however, that retrieval will in general be secondary to the publication of journals, abstracts and indexes. Again, it is important that the flexibility of computer typesetting be exploited fully, especially by providing bibliographic tools for human use in as wide a variety of forms as possible[23]. Recent work in Sheffield, for example, has indicated that automation can assist the compilers of indexes still further; articulated subject indexes, similar in structure to those of *Chemical Abstracts*, can be constructed by computer from simple title-like phrases chosen by human indexers, thus relieving indexers and editors of much repetitive work[24].

The storage and manipulation of information on chemical structures present a strong contrast to the problems involved in dealing with information expressed in linguistic terms. Topological descriptions of molecules are ideally suited to computer organization and search. Today, general methods exist for assigning unique machine descriptions to discrete chemical structures[25], and for searching molecules for specified substructures[26–28]. In particular, the substructure searches free the searcher from constraints imposed by hierarchies which are implicit in chemical nomenclatures. Because more than 3·5 million different chemical compounds are already known, and the number increases by about 75,000 each year, it is essential to use computers to organize this information efficiently. The Chemical Abstracts Service is at present engaged in a project which will lead to a file in which every published compound will be included; today, almost 500,000 structures are available in machine form. This file is being used to aid in naming compounds for inclusion in manual indexes, and will eventually be used to provide structure search services[29]. The potentialities of these searches have already evoked widespread interest among chemists. Studies which would extend the range of manipulations possible on chemical structures are at present in hand at Sheffield, where computer techniques to determine similarities among structures are being developed[30]. These differ from the straightforward search techniques in that a structure or substructure need not be specified at the outset; rather, the fragment common to a pair of structures is determined automatic-

ally. The method could thus be used to analyse structural changes in chemical reactions, or, eventually, to discover correlations between structures and activities.

The searches at present provided by the information services referred to at the outset differ considerably from the procedures that a scientist uses when he consults an index or other source, where new lines of pursuit may suggest themselves constantly, and may be modified according to the number and nature of the references discovered. The scope for interactions of this sort is limited with present machine searches. A query is submitted, and the results are returned some hours or days later. The search request can be improved and the results only evaluated over a considerable period of time. There are strong indications that the use of multiple access computers, which a number of users can interrogate virtually simultaneously, may fill the gap caused by the slow response time associated with batch processing. A multiple access machine can have a number of typewriters linked to it, possibly from remote locations. The user can type in his query, and has an almost instantaneous response typed out before him. He can refine the terms of his search, narrowing it if the output is large, widening it if small. (He may also request guidance, by calling for lists of words associated with those of his initial enquiry to be typed out.) This approach is already a reality, if to a limited extent at present, with Project MAC at the Massachusetts Institute of Technology, where small files of information on physics can be searched by scientists who have teletypewriters in their offices, laboratories and even in their homes. This is closely allied to the ambitious Intrex project (Information Transfer Experiment)[31], the purpose of which is to use existing technology to exploit the resources of a university library to best advantage. Building on material already available in digital form (MEDLARS and NASA data), the project will make it possible for the scientist to search complete catalogues of the library's holdings at his own desk, thus reducing the effort he must expend in order to gain access to information. In this project, which owes much to Mooers's promotion of the concept of the "reactive typewriter"[32], a battery of devices which will provide a variety of levels of access to the collection is foreseen. The telephone is the simplest of these; the computer, when interrogated by dialling, could provide recorded replies to a limited range of directory enquiries. A teletypewriter could also be used, and would provide information at a greater rate, and with more scope for interaction. At the third level, a television screen could display information at a very fast rate, when linked to the computer by a keyboard. Rapid copying and transmission of documents are also foreseen in the scope of the experiment, which should give valuable insight into the interactions between men and automated information systems.

Decentralization

It seems certain that the ultimate success of information services depends not only on the adequacy of the actual retrieval process but also on the means by which the services and their users are linked. Thus, it is interesting to note that a trend towards decentralization has already begun. The U.S. National Library of Medicine has already acted in this direction by establishing secondary centres in California, Colorado, Michigan and New England, as well as in Britain and Sweden. The Intrex project carries this trend yet further, to the individuals at their desks. In the light of these developments, it is to be hoped that university libraries in Britain will be given the staff, equipment and funds to enable them to provide academic staff with the active information services that their colleagues in industrial and governmental research establishments have come to regard as their right, and that this will advance hand in hand with the increasing use of computers to control the housekeeping activities of libraries[33].

The channels for communicating information provided by the scientific societies and by commercial and governmental organizations are important, yet they form only a part of national and international networks, of the workings of which we have as yet only an incomplete picture. In these networks, primary journal publication, abstracting and indexing services, preprint exchange, citation of previous work, conferences, colloquia and personal exchanges all have their place, each characterized by a different time scale. The channels for exchange of information are dynamic systems, which change constantly in response to changing needs. Timeliness is one of the most important factors which influence the changes, and the need for it has undoubtedly led to the recent growth of informal routes such as preprint exchange. Not only the traditional forms of publication, but also those which bypass the delays associated with formal publication, stand in need of study, and of support, if their role can be shown to be useful[34]. Furthermore, it is important to be able to predict the effects of changes in one area on activities in others. It seems possible that an approach using the methods of cybernetics could aid in the description and understanding of the significant interactions in the exchange of information on a national or international scale. It could lead to a simulation of the information transfer system, analogous to the studies of the national economy in progress at the University of Cambridge under Stone[35], and could give insight into the dynamic aspects of the system which could help guide the efforts of the administrators who must ensure improved flow and utilization of information. The problems which would beset such an approach are manifold; not the least of them is the difficulty of measuring the variables involved.

In the final analysis, it is the scientists and engineers themselves who must be the arbiters of the usefulness of innovations in information handling. Without their active participation and criticism at this early and imperfect stage, progress towards the most effective use of computers in the communication of scientific information must be slow.

[1] *Bull. Med. Lib. Assoc.*, **52**, 148 (1964).
[2] Freeman, R. R., and Dyson, G. M., *J. Chem. Document.*, **3**, 16 (1963).
[3] Dyson, G. M., and Lynch, M. F., *J. Chem. Document.*, **3**, 81 (1963).
[4] Zabriskie, K. H., and Lynch, M. F., *J. Chem. Document.*, **6**, 30 (1966).
[5] Aitchison, T. M., *Aslib Proc.*, **17**, 343 (1965).
[6] Luhn, H. P., *Amer. Document.*, **12**, 131 (1961).
[7] Luhn, H. P., *Amer. Document.*, **11**, 288 (1960).
[8] Metcalfe, J. W., *Information Indexing and Subject Cataloging* (Scarecrow Press, New York, 1957).
[9] Cleverdon, C., Mills, J., and Keen, M., *Factors Determining the Performance of Indexing Systems* (Aslib Cranfield Research Project, 1966).
[10] Luhn, H. P., *IBM J. Res. Dev.*, **1**, 309 (1957).
[11] Baxendale, P. B., *IBM J. Res. Dev.*, **2**, 354 (1958).
[12] Maron, M. E., and Kuhns, J. L., *J. Assoc. Comp. Mach.*, **7**, 216 (1960).
[13] Maron, M. E., *J. Assoc. Comp. Mach.*, **8**, 404 (1961).
[14] Doyle, L. B., *J. Assoc. Comp. Mach.*, **8**, 553 (1961).
[15] Stiles, H. E., *J. Assoc. Comp. Mach.*, **8**, 271 (1961).
[16] Needham, R. M., in *Information Processing 1962*, 284 (1962).
[17] Meetham, A. R., *Language and Speech*, **6**, 22 (1963).
[18] Vaswani, P., *Rev. Intern. Document.*, **32**, 19 (1965).
[19] Giuliano, V. E., *IRE Trans. Mil. Electron., MIL-7*, 221 (1963).
[20] Salton, G., *Commun. Assoc. Comp. Mach.*, **9**, 204 (1966).
[21] Gardin, J.-C., *Syntol* (*Syntagmatic Organisation Language*) (The State University, Rutgers, New Jersey, 1963).
[22] Meetham, A. R., Paper presented at Assoc. Comp. Mach. Ann. Mtg., August 1966.
[23] Coblans, H., *Use of Mechanised Methods in Documentation Work* (Aslib, London, 1966).
[24] Lynch, M. F., *J. Document.*, **22**, 167 (1966).
[25] Morgan, H. L., *J. Chem. Document.*, **5**, 107 (1965).
[26] Cossum, W. E., Krakiwsky, M. L., and Lynch, M. F., *J. Chem. Document.* **5**, 33 (1965).
[27] Sussenguth, E. H., *J. Chem. Document.*, **5**, 36 (1965).
[28] Gould, D., Gasser, E. B., and Rian, J. F., *J. Chem. Document.*, **5**, 24 (1965).
[29] *Annual Report to the National Science Foundation* (Chemical Abstracts Service, Columbus, Ohio, 1966).
[30] Armitage, J. E., and Lynch, M. F., *J. Chem. Soc. Org.* (in the press).
[31] Overhage, C. F. J., *Science*, **152**, 1032 (1966).
[32] Mooers, C. N., *Commun. Assoc. Comp. Mach.*, **9**, 215 (1966).
[33] Line, M. B., *Brit. Univ. Ann.*, 93 (1965).
[34] *Nature*, **212**, 867 (1966).
[35] Stone, R., *Operat. Res. Quart.*, **14**, 51 (1963).